A finite volume method on general meshes for a degenerate parabolic convection-reaction-diffusion equation.

Ophélie Angelini, Konstantin Brenner, Danielle Hilhorst

A finite volume method on general meshes for a degenerate parabolic convection-reaction-diffusion equation

Ophlie Angelini [*], Konstantin Brenner [†], Danielle Hilhorst [‡]

November 17, 2010

Abstract We propose a finite volume method on general meshes for the discretization of a degenerate parabolic convection-reaction-diffusion equation. Equations of this type arise in many contexts, such as the modeling of contaminant transport in porous media. We discretize the diffusion term, which can be anisotropic and heterogeneous, via a hybrid finite volume scheme. We construct a partially upwind scheme for the convection term. We consider a wide range of unstructured possibly non-matching polygonal meshes in arbitrary space dimension. The only assumption on the mesh is that the volume elements must be star-shaped. The scheme is fully implicit in time, it is locally conservative and robust with respect to the Pclet number. We obtain a convergence result based upon a priori estimates and the Fréchet–Kolmogorov compactness theorem.

1   Introduction

In this paper we study a finite volume method on general meshes for degenerate parabolic convection-reaction-diffusion equations of the form

$$\frac{\partial \beta(u)}{\partial t} - \nabla \cdot (\boldsymbol{\Lambda} \nabla u) + \nabla \cdot (\mathbf{V} u) + F(u) = q. \tag{1}$$

Equations of this type arise in particular in the modeling of contaminant transport in groundwater. The unknown function u represents the concentration of the species, which diffuses and is transported by the groundwater. An essential element in our study is the processus of adsorbtion by a porous skeleton, which is supposed to be very fast. More particularly we

---

[*]EDF R&D, 1 avenue du Gnral de Gaulle 92141 Clamart, France

[†]Laboratoire de Mathématiques, Université de Paris-Sud 11, F-91405 Orsay Cedex, France

[‡]CNRS and Laboratoire de Mathématiques, Université de Paris-Sud 11, F-91405 Orsay Cedex, France

suppose that the dissolved and the absorbed parts of the species are in equilibrium; this is modeled by the function β, where β' may be infinite in several points. The matrix $\mathbf{\Lambda}$ is a possibly anisotropic and heterogeneous diffusion-dispersion tensor, $\mathbf{V}$ is the velocity field, the function F stands for the chemical reactions, and q is the source term. We suppose that the mesh is quite general, and possibly nonmatching. Therefore, also in view of the anisotropy in the diffusion term, we can not apply the standard finite volume method [13].

Finite volume schemes have often been applied to the equation (1), see e.g. [2], [6], [15]. The upwind discretization of the convection term permits finite volume schemes to be stable in convection dominated case, however standard finite volume schemes do not permit to handle anisotropic diffusion on general meshes. On the other hand finite element method allows a very simple discretization of full diffusion tensors, they were used a lot for the discretization of equation (1), see e.g. [5], [9], [10]. A possible solution is to split equation (1) into a hyperbolic part and a parabolic part, by means of an operator splitting method; one can find such an analysis in [21], [22], where the advection term was treated by the method of characteristics. The other quite intuitive idea is to take "best from both worlds" [17], which leads to combined finite volume-finite element schemes; we refer to [17] for this approach. In order to solve this class of equations, Eymard, Hilhorst and Vohralík [17] discretize the diffusion term by means of piecewise linear nonconforming (Crouzeix–Raviart) finite elements over a triangularization of the space domain, or using the stiffness matrix of the hybridization of the lowest order Raviart–Thomas mixed finite element method. The other terms are discretized by means of a finite volume scheme on a dual mesh, where the dual volumes are constructed around the sides of the original triangularization. In the second paper of Eymard et al. [18] the time evolution, convection, reaction, and sources terms are discretized on a given grid, which can be nonmatching and can contain nonconvex elements, by means of a cell-centered finite volume method. In order to discretize the diffusion term, they construct a conforming simplicial mesh with vertices given by the original grid and use the finite element method. In this way, the scheme is fully consistent and the discrete solution is naturally continuous across the interfaces between the subdomains with nonmatching grids, without introducing supplementary equations and unknowns nor interpolating the discrete solutions at the interfaces.

The finite volume methods for the discretization of anisotropic diffusion on general meshes is a subject of wide interest (see for instance the results of the benchmarks organized at the FVCA 5 conference [FVCA5]). We refer to [12] for a detailed analysis of three recently developed families of schemes, namely the Mimetic Finite Difference scheme, the Hybrid Finite Volume scheme and the Mixed Finite Volume, which turn out to be quite similar. The most important feature of these methods is their accurate approximation of anisotropic diffusion even in highly heterogenous cases. In this paper, we apply a recent method based upon the finite volume method on general meshes developed by Eymard, Gallouët et Herbin [13], whereas we use a slightly modified upwind scheme for the approximation of the convection term. The time discretization is based upon a completely implicit finite difference scheme.

The organization of this paper is as follows. We describe the numerical scheme in Section 2. We show the existence and uniqueness of the solution of the discrete scheme and prove a priori

estimates for the discrete solution in $L^\infty(0, T; L^2(\Omega))$ and in a discrete space analogous to the space $L^2(0, T; H^1(\Omega))$ in Section 3. In Section 4, we prove an estimate on differences of time translates whereas we establish an estimate on differences of space translates in Section 5. These estimates imply a relative compactness property of sequences of approximate solutions by the Fréchet–Kolmogorov theorem. We deduce the strong convergence in L^2 of the approximate solutions to the unique solution of the continuous problem in Section 6. For the proofs, we apply methods inspired upon those of [13] and [14]. In Section 7, we finally present results of numerical tests, which confirm the validity of the numerical method.

2 The numerical scheme

We consider the parabolic degenerate convection-diffusion-reaction problem

$$(\mathcal{P}) \begin{cases} \dfrac{\partial \beta(u)}{\partial t} - \nabla \cdot (\mathbf{\Lambda}(\mathbf{x})\nabla u) + \nabla \cdot (\mathbf{V}(\mathbf{x})u) + F(u) = q(\mathbf{x}, t), & (\mathbf{x}, t) \in Q_T, \\[2mm] u(\mathbf{x}, t) = 0 \quad \mathbf{x} \in \partial\Omega, \quad t \in (0, T), \\[2mm] u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \end{cases}$$

where Ω is a bounded open connected polyhedral subset of $\mathbb{R}^d$, $d \in \mathbb{N} \setminus \{0\}$, $T > 0$ and $Q_T = \Omega \times (0, T)$. We suppose that the following hypotheses are satisfied:

$(\mathcal{H}_1)$ $\beta \in C(\mathbb{R}), \beta(0) = 0$ is a strictly increasing function, which satisfies the growth condition $|\beta(a) - \beta(b)| \geqslant \underline{\beta}|a - b|$, $\underline{\beta} > 0$ for all $a, b \in \mathbb{R}$; moreover there exist $P > 0$ and C_β, such that $|\beta(u)| \leqslant C_\beta$ for $|u| \leqslant P$ and β is a Lipschitz continuous with a constant $\overline{\beta}$ for $|u| \geqslant P$;
$(\mathcal{H}_2)$ $\mathbf{\Lambda}$ is a measurable function from Ω to $\mathcal{M}_d(\mathbb{R})$, where $\mathcal{M}_d(\mathbb{R})$ denotes the set of $d \times d$ symmetric matrices, such that for a.e. $\mathbf{x} \in \Omega$ the set of its eigenvalues is included in $[\lambda_m, \lambda_M]$, where $\lambda_m, \lambda_M \in L^\infty(\Omega)$ are such that $0 < \underline{\lambda} \leqslant \lambda_m(\mathbf{x}) \leqslant \lambda_M(\mathbf{x}) \leqslant \overline{\lambda}$;
$(\mathcal{H}_3)$ $\mathbf{V} \in H(\mathrm{div}, \Omega) \bigcap \mathrm{L}^\infty(\Omega)$ is such that $\nabla \cdot \mathbf{V} \geqslant 0$ a.e. in Ω;
$(\mathcal{H}_4)$ $u_0 \in L^\infty(\Omega)$;
$(\mathcal{H}_5)$ $F \in C(\mathbb{R})$, $F(0) = 0$ and there exists $M > 0$ such that $uF(u) > 0$ and $F(u)$ is Lipschitz continuous with constant L_F for all $u < 0$ or $u > M$; moreover we suppose that F does not decrease too fast i.e. there exists $\underline{F} > 0$ such that $(F(u) - F(v))(u - v) \geqslant -\underline{F}(u - v)^2$ for all $u, v \in \mathbb{R}$;
$(\mathcal{H}_6)$ $q \in L^2(Q_T)$.
We now present a definition of a weak solution of Problem $(\mathcal{P})$.

Definition 2.1. *We say that a function u is a weak solution of Problem $(\mathcal{P})$ if*
(i) $u \in L^2(0, T; H_0^1(\Omega))$;
(ii) $\beta(u) \in L^\infty(0, T; L^2(\Omega))$;
(iii) u *satisfies the integral equality*

$$-\int_0^T \int_\Omega \beta(u)\varphi_t \, d\mathbf{x}dt - \int_\Omega \beta(u_0)\varphi(\cdot, 0) \, d\mathbf{x} + \int_0^T \int_\Omega \mathbf{\Lambda}\nabla u \cdot \nabla\varphi \, d\mathbf{x}dt$$

3

$$- \int_0^T \int_\Omega u\mathbf{V} \cdot \nabla\varphi \, d\mathbf{x}dt + \int_0^T \int_\Omega F(u)\varphi \, d\mathbf{x}dt = \int_0^T \int_\Omega q\varphi \, d\mathbf{x}dt$$

for all $\varphi \in L^2(0,T; H_0^1(\Omega))$ with $\varphi_t \in L^\infty(Q_T)$, $\varphi(\cdot, T) = 0$.

Remark 2.1. *In the case that the reaction function F is nondecreasing, the uniqueness of the weak solution of Problem $(\mathcal{P})$ follows from [24].*

In order to describe the numerical scheme we introduce below some notations related to the space and time discretization.

Definition 2.2. (Space discretization) *Let Ω be a polyhedral open bounded connected subset of $\mathbb{R}^d$, with $d \in \mathbb{N} \setminus \{0\}$, and $\partial\Omega = \overline{\Omega}\backslash\Omega$ its boundary. A discretization of Ω, denoted by $\mathcal{D}$, is defined as the triplet $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{P})$, where:*

1. $\mathcal{M}$ is a finite family of non empty connex open disjoint subsets of Ω (the "control volumes") such that $\bar{\Omega} = \bigcup_{K \in \mathcal{M}} K$. For any $K \in \mathcal{M}$, let $\partial K = \overline{K}\backslash K$ be the boundary of K; we define $m(K) > 0$ as the measure of K and h_K as the diameter of K.

2. $\mathcal{E}$ is a finite family of disjoint subsets of $\overline{\Omega}$ (the "edges" of the mesh), such that, for all $\sigma \in \mathcal{E}$, σ is a non empty open subset of a hyperplane of $\mathbb{R}^d$, whose $(d-1)$-dimensional measure $m(\sigma)$ is strictly positive. We also assume that, for all $K \in \mathcal{M}$, there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \bigcup_{\sigma \in \mathcal{E}_K} \sigma$. For each $\sigma \in \mathcal{E}$, we set $\mathcal{M}_\sigma = \{K \in \mathcal{M} | \sigma \in \mathcal{E}_K\}$. We then assume that, for all $\sigma \in \mathcal{E}$, either $\mathcal{M}_\sigma$ has exactly one element and then $\sigma \in \partial\Omega$ (the set of these interfaces called boundary interfaces, is denoted by $\mathcal{E}_{ext}$) or $\mathcal{M}_\sigma$ has exactly two elements (the set of these interfaces called interior interfaces, is denoted by $\mathcal{E}_{int}$). For all $\sigma \in \mathcal{E}$, we denote by $\mathbf{x}_\sigma$ the barycenter of σ. For all $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}_K$, we denote by $\mathbf{n}_{K,\sigma}$ the outward normal unit vector.

3. $\mathcal{P}$ is a family of points of Ω indexed by $\mathcal{M}$, denoted by $\mathcal{P} = (\mathbf{x}_K)_{K \in \mathcal{M}}$, such that for all $K \in \mathcal{M}$, $\mathbf{x}_K \in K$; moreover K is assumed to be $\mathbf{x}_K$-star-shaped, which means that for all $\mathbf{x} \in K$, there holds $[\mathbf{x}_K, \mathbf{x}] \in K$. Denoting by $d_{K,\sigma}$ the Euclidean distance between $\mathbf{x}_K$ and the hyperplane containing σ, one assumes that $d_{K,\sigma} > 0$. We denote by $D_{K,\sigma}$ the cone of vertex $\mathbf{x}_K$ and basis σ.

Next we introduce some extra notations related to the mesh. The size of the discretization $\mathcal{D}$ is defined by

$$h_\mathcal{D} = \sup_{K \in \mathcal{M}} diam(K); \tag{2}$$

moreover we define

$$\theta_\mathcal{D} = \max(\max_{\sigma \in \mathcal{E}_{int}, \{K,L\}=\mathcal{M}_\sigma} \frac{d_{K,\sigma}}{d_{L,\sigma}}, \max_{K \in \mathcal{M}_\sigma, \sigma \in \mathcal{E}_K} \frac{h_K}{d_{K,\sigma}}). \tag{3}$$

Thus imposing a uniform bound on $\theta_\mathcal{D}$ forces the meshes to be sufficiently regular. As it was done in [14] we associate with the mesh the following spaces of discrete unknowns

$$\begin{aligned} X_\mathcal{D} &= \{((v_K)_{K \in \mathcal{M}}, (v_\sigma)_{\sigma \in \mathcal{E}}), v_K \in \mathbb{R}, v_\sigma \in \mathbb{R}\}, \\ X_{\mathcal{D},0} &= \{v \in X_\mathcal{D} \text{ such that } (v_\sigma)_{\sigma \in \mathcal{E}_{ext}} = 0\}. \end{aligned} \tag{4}$$

Moreover, for each function $\varphi = \varphi(\mathbf{x})$ smooth enough we define $P_{\mathcal{D}}\varphi \in X_D$ in following way

$$(P_{\mathcal{D}}\varphi)_K = \varphi(\mathbf{x}_K) \quad \text{for all} \quad K \in \mathcal{M},$$
$$(P_{\mathcal{D}}\varphi)_\sigma = \varphi(\mathbf{x}_\sigma) \quad \text{for all} \quad \sigma \in \mathcal{E}.$$

Definition 2.3. (Time discretization) *We divide the time interval $(0,T)$ into N equal time steps of length $\delta t = T/N$ such that*

$$\delta t < \underline{\beta}/\underline{F}, \tag{5}$$

where δt is the uniform time step defined by $\delta t = t_n - t_{n-1}$.

Remark 2.2. *For the sake of simplicity, we restrict our study to the case of constant time steps. Nevertheless all results presented below can be easily extended to the case of a non uniform time discretization.*

After formally integrating the first equation of $(\mathcal{P})$ on the domain $K \times (t_{n-1}, t_n)$ for each $K \in \mathcal{M}$ and $n = 1, \ldots, N$, we obtain

$$\int_K \beta(u(\mathbf{x}, t_n)) - \beta(u(\mathbf{x}, t^{n-1}))\ d\mathbf{x} + \sum_{\sigma \in \mathcal{E}_K} \int_{t_{n-1}}^{t_n} \int_\sigma (-\mathbf{\Lambda}\nabla u + \mathbf{V}u) \cdot \mathbf{n}_{K,\sigma}\ d\gamma dt$$
$$+ \int_{t_{n-1}}^{t_n} \int_K F(u)\ d\mathbf{x}dt = \int_{t_{n-1}}^{t_n} \int_K q\ d\mathbf{x}dt.$$

For all $K \in \mathcal{M}$ and all $\sigma \in \mathcal{E}_K$ we define $V_{K,\sigma} = \int_\sigma \mathbf{V} \cdot \mathbf{n}_{K,\sigma}d\gamma$ and $q_K^n = \dfrac{1}{\delta t\ m(K)} \int_{t_{n-1}}^{t_n} \int_K q\ d\mathbf{x}dt$. We use an upwind scheme in order to approximate the convective term, since it can possibly dominate the diffusion term; the diffusive flux $-\int_\sigma \mathbf{\Lambda}\nabla u \cdot \mathbf{n}_{K,\sigma}d\gamma$ is approximated by a function of the form $F_{K,\sigma}(u^n)$, where $u^n = ((u_K^n)_{K\in\mathcal{M}}, (u_\sigma^n)_{\sigma\in\mathcal{E}})$, and where the numerical flux $F_{K,\sigma}(u^n)$ is defined by formula (25) below. The time implicit finite volume scheme corresponding to Problem $(\mathcal{P})$ is given by:

(i) The initial condition

$$u_K^0 = \frac{1}{m(K)} \int_K u_0(\mathbf{x})\ d\mathbf{x}, \tag{6}$$

for all $K \in \mathcal{M}$.

(ii) The discrete equations

$$m(K)(\beta(u_K^n) - \beta(u_K^{n-1})) + \delta t \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}(u^n) + \delta t \sum_{\sigma \in \mathcal{E}_K} V_{K,\sigma}\overline{u_{K,\sigma}^n}$$
$$+ \delta t\ m(K)F(u_K^n) = \delta t\ m(K)q_K^n, \tag{7}$$

for all $K \in \mathcal{M}$. Unlike in the case of the standard upwind scheme, we define the value $\overline{u_{K,\sigma}^n}$ as follows. For all $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}_K$ we set

$$\overline{u_{K,\sigma}^n} = \begin{cases} u_K^n, & \text{if } V_{K,\sigma} \geqslant 0 \\ u_\sigma^n, & \text{if } V_{K,\sigma} < 0. \end{cases} \tag{8}$$

5

We also define

$$V_{K,\sigma}^+ = \frac{1}{2}(V_{K,\sigma} + |V_{K,\sigma}|) \quad \text{and} \quad V_{K,\sigma}^- = \frac{1}{2}(V_{K,\sigma} - |V_{K,\sigma}|) \tag{9}$$

which lead to

$$V_{K,\sigma}\overline{u_{K,\sigma}^n} = V_{K,\sigma}^+ u_K^n + V_{K,\sigma}^- u_\sigma^n \tag{10}$$

The definition of (8) seems natural since we also take the unknowns associated with the mesh faces. It has an important advantage that the unknowns in the equation (7) are associated with a single control volume (see Remark 7.1); moreover the numerical experiments presented in Section 7 show that the upwind scheme (8) also preserves the approximate solution from unphysical oscillations in the convection dominated case. Finally, we remark that for each time step the number of equations is $card(\mathcal{M})$, whereas the number of discrete unknowns is equal to $card(\mathcal{M}) + card(\mathcal{E})$. Therefore we need to introduce $card(\mathcal{E})$ additional equations corresponding to the interface values. For boundary faces these equations are obtained by writing the discrete analog of the Dirichlet boundary condition

$$(\text{iii}) \quad u_\sigma^n = 0 \quad \text{for all} \quad \sigma \in \mathcal{E}_{ext}. \tag{11}$$

For interior faces, we follow the main idea of the finite volume method by imposing the local conservation of the discrete fluxes

$$(\text{iv}) \quad (F_{K,\sigma}(u^n) + V_{K,\sigma}\overline{u_{K,\sigma}^n}) + (F_{L,\sigma}(u^n) + V_{L,\sigma}\overline{u_{L,\sigma}^n}) = 0 \tag{12}$$

for all $\sigma \in \mathcal{E}_{int}$ with $\mathcal{M}_\sigma = \{K, L\}$. We will define below $F_{K,\sigma}$ in some more detail, but we first give an alternative variational formulation of the discrete scheme (i)-(iv). Let $\{v^n\}_{n\in\mathbb{N}}$ be an arbitrary sequence of elements of $X_{\mathcal{D},0}$; multiplying equation (7) by v_K^n and summing on all control volumes $K \in \mathcal{M}$ leads to:

$$\sum_{K\in\mathcal{M}} m(K)v_K^n \frac{\beta(u_K^n) - \beta(u_K^{n-1})}{\delta t} + \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} (v_K^n F_{K,\sigma}(u^n) + v_K^n V_{K,\sigma}\overline{u_{K,\sigma}^n})$$
$$+ \sum_{K\in\mathcal{M}} m(K)v_K^n F(u_K^n) = \sum_{K\in\mathcal{M}} m(K)v_K^n q_K^n.$$

Using (12), we obtain that

$$\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} v_\sigma^n(F_{K,\sigma}(u^n) + V_{K,\sigma}\overline{u_{K,\sigma}^n}) = 0 \text{ for all } v^n \in X_{\mathcal{D}}, \tag{13}$$

which yields the following discrete weak formulation:

Let u_K^0 be defined by:

$$u_K^0 = \frac{1}{m(K)} \int_K u_0(\mathbf{x})\, d\mathbf{x} \quad \text{for all } K \in \mathcal{M} \tag{14}$$

For each $n \in \{1, \ldots, N\}$ find $u^n \in X_{\mathcal{D},0}$ such that for all $v^n \in X_{\mathcal{D},0}$:

$$\sum_{K \in \mathcal{M}} m(K) v_K^n \frac{\beta(u_K^n) - \beta(u_K^{n-1})}{\delta t} + < v^n, u^n >_F + < v^n, u^n >_T$$
$$+ \sum_{K \in \mathcal{M}} m(K) v_K^n F(u_K^n) = \sum_{K \in \mathcal{M}} m(K) v_K^n q_K^n, \tag{15}$$

with

$$< v, u >_F = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} (v_K - v_\sigma) F_{K,\sigma}(u) \tag{16}$$

and

$$< v, u >_T = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} (v_K - v_\sigma) V_{K,\sigma} \overline{u_{K,\sigma}}. \tag{17}$$

Remark that the problem (i)-(iv) is equivalent to (15)-(17). Indeed, let δ_{ij} be the Kroneker symbol, by setting $v_\sigma^n = 0$ for all $\sigma \in \mathcal{E}$, and $v_{K'}' = \delta_{KK'}$ for all $K' \in \mathcal{M}$ and for a given K one recover (ii), and setting $v_K = 0$ for all $K \in \mathcal{M}$ and $v_{\sigma'} = \delta_{\sigma\sigma'}$ for all $\sigma' \in \mathcal{E}$ yields (iv). The homogeneous Dirichlet boundary condition (iii) follows from the fact that $u^n \in X_{\mathcal{D},0}$. In order to complete the numerical scheme we still have to express the discrete flux $F_{K,\sigma}$ in terms of the discrete unknowns. For this purpose we use the SUSHI scheme proposed in [14]: the idea is based upon the identification of the numerical fluxes $F_{K,\sigma}$ through the mesh dependent bilinear form, using the expression of a discrete gradient. We first define

$$\nabla_K u = \frac{1}{m(K)} \sum_{\sigma \in \mathcal{E}_K} m(\sigma)(u_\sigma - u_K) \mathbf{n}_{K,\sigma} \quad \forall K \in \mathcal{M}, \quad \forall u \in X_{\mathcal{D}}. \tag{18}$$

Remark that the geometrical relation

$$\sum_{\sigma \in \mathcal{E}_K} m(\sigma) \mathbf{n}_{K,\sigma} (\mathbf{x}_\sigma - \mathbf{x}_K)^T = m(K) Id \tag{19}$$

holds for each K. Let $\varphi(\mathbf{x})$ be a function, piecewise linear on the control volumes of the mesh. In view of (19) one has $\nabla_K P_{\mathcal{D}}(\varphi) = \nabla\varphi(\mathbf{x})|_{x \in K}$. We also remark that

$$\sum_{\sigma \in \mathcal{E}_K} m(\sigma) \mathbf{n}_{K,\sigma} = \sum_{\sigma \in \mathcal{E}_K} \int_\sigma \mathbf{n}_{K,\sigma} \, d\gamma = \int_K \nabla 1 \, d\mathbf{x} = 0,$$

which means that the coefficient of u_K in (18) is equal to zero; thus, a reconstruction of the discrete gradient solely based on (18) cannot lead to a coercive discrete bilinear form in the general case. Therefore we introduce the additional term

$$\nabla_{K,\sigma} u = \nabla_K u + R_{K,\sigma} u \cdot \mathbf{n}_{K,\sigma}, \tag{20}$$

where

$$R_{K,\sigma} u = \frac{\alpha_K}{d_{K,\sigma}} (u_\sigma - u_K - \nabla_K u \cdot (\mathbf{x}_\sigma - \mathbf{x}_K)), \tag{21}$$

for some $\alpha_K > 0$, which should be chosen in a suitable way. If we choose $\alpha_K = \sqrt{d}$ for all $K \in \mathcal{M}$ in the simple case that Λ is a scalar and that the mesh that satisfies the orthogonality property $\mathbf{n}_{K,\sigma} = \dfrac{\mathbf{x}_\sigma - \mathbf{x}_K}{d_{K,\sigma}}$, we obtain the usual two point scheme. Nevertheless, it may be useful to optimize the choice of α_K as it is done in [3]. We then define the discrete gradient $\nabla_\mathcal{D} u$ as the piecewise constant function equal to $\nabla_{K,\sigma} u$ in the cone $D_{K,\sigma}$ with vertex $\mathbf{x}_K$ and basis σ

$$\nabla_\mathcal{D} u|_{D_{K,\sigma}} = \nabla_{K,\sigma} u.$$

Note that the term $R_{K,\sigma}$ is a second order error term, which vanishes for piecewise linear functions. Moreover, the relation (19) together with (21) implies that

$$\sum_{\sigma \in \mathcal{E}_K} m(D_{K,\sigma}) R_{K,\sigma}(u) \mathbf{n}_{K,\sigma} = 0 \text{ for all } K \in \mathcal{M} \text{ and for all } u \in X_\mathcal{D}, \tag{22}$$

which in turn implies that

$$\int_K \nabla_\mathcal{D} u \, d\mathbf{x} = m(K) \nabla_K u.$$

The discrete gradient defined above satisfies the following strong consistency property.

Lemma 2.1. *Let $\mathcal{D}$ be a discretization of Ω in sense of Definition 2.2, moreover let $\theta \geqslant \theta_\mathcal{D}$ be given. Then for all $\varphi \in C^2(\overline{\Omega})$, there exist a positive constant C only depending on d, θ and φ such that*

$$\|\nabla_\mathcal{D} P_\mathcal{D} \varphi - \nabla \varphi\|_{(L^\infty(\Omega))^d} \leqslant Ch_\mathcal{D}.$$

The proof of this Lemma is given in [14]. The numerical flux is implicitly defined by the relation

$$< v, u >_F = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} (v_K - v_\sigma) F_{K,\sigma}(u) = \int_\Omega \nabla_\mathcal{D} v \cdot \Lambda(\mathbf{x}) \nabla_\mathcal{D} u \, d\mathbf{x}. \tag{23}$$

It can also be defined explicitly; in order to do so, we write the discrete gradient in the form

$$\nabla_{K,\sigma} u = \sum_{\sigma' \in \mathcal{E}_K} (u_{\sigma'} - u_K) \mathbf{y}^{\sigma\sigma'}, \tag{24}$$

where $\mathbf{y}^{\sigma\sigma'}$ is defined by

$$\mathbf{y}^{\sigma\sigma'} = \begin{cases} \dfrac{m(\sigma)}{m(K)} \mathbf{n}_{K,\sigma} + \dfrac{\sqrt{d}}{d_{K,\sigma}} (1 - \dfrac{m(\sigma)}{m(K)} \mathbf{n}_{K,\sigma} \cdot (\mathbf{x}_\sigma - \mathbf{x}_K)) \mathbf{n}_{K,\sigma} & \text{if } \sigma = \sigma', \\ \dfrac{m(\sigma')}{m(K)} \mathbf{n}_{K,\sigma'} - \dfrac{\sqrt{d}}{d_{K,\sigma}} \dfrac{m(\sigma')}{m(K)} \mathbf{n}_{K,\sigma'} \cdot (\mathbf{x}_\sigma - \mathbf{x}_K) \mathbf{n}_{K,\sigma} & \text{otherwise.} \end{cases}$$

We obtain that for all $u, v \in X_\mathcal{D}$

$$\int_\Omega \nabla_\mathcal{D} u(\mathbf{x}) \cdot \Lambda(\mathbf{x}) \nabla_\mathcal{D} v(\mathbf{x}) = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} \sum_{\sigma' \in \mathcal{E}_K} A_K^{\sigma\sigma'} (u_\sigma - u_K)(v_{\sigma'} - v_K),$$

with

$$A_K^{\sigma\sigma'} = \sum_{\sigma'' \in \mathcal{E}_K} \mathbf{y}^{\sigma''\sigma} \cdot \mathbf{\Lambda}_{K,\sigma''} \mathbf{y}^{\sigma''\sigma'} \text{ and } \mathbf{\Lambda}_{K,\sigma''} = \int_{D_{K,\sigma''}} \mathbf{\Lambda}(\mathbf{x}) \, d\mathbf{x}.$$

The local matrices $(A_K^{\sigma\sigma'})_{\sigma\sigma' \in \mathcal{E}_K}$ are symmetric, and the numerical flux is then defined by

$$F_{K,\sigma}(u) = \sum_{\sigma' \in \mathcal{E}_K} A_K^{\sigma\sigma'}(u_K - u_{\sigma'}). \tag{25}$$

Next we prove some useful properties of the mesh depending bilinear forms introduced previously. In particular we prove that $< \cdot, \cdot >_F$ is continuous and coercive, and that $< \cdot, \cdot >_T$ is continuous and nonnegative. The space $X_\mathcal{D}$ defined in (4) is equipped with the following semi-norm.

Definition 2.4. *Let $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{P})$ be a discretization of Ω in the sense of Definition 2.2; then for all $v \in X_\mathcal{D}$ we define*

$$|v|_X^2 = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} \frac{m(\sigma)}{d_{K,\sigma}}(v_\sigma - v_K)^2, \tag{26}$$

which is a norm on the space $X_{\mathcal{D},0}$. Let us also define a discrete analog of $\| \cdot \|_{1,p}$ norm.

Definition 2.5. (The discrete space $H_\mathcal{M}(\Omega)$) *Let $1 \leqslant p < \infty$ and let $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{P})$ be a discretization of Ω in the sense of Definition 2.2. Let $H_\mathcal{M}(\Omega) \subset L^2(\Omega)$ be the set of piecewise constant functions on the control volumes of the mesh $\mathcal{M}$ for each $v \in H_\mathcal{M}(\Omega)$ we define $v_K = v(\mathbf{x})|_{\mathbf{x} \in K}$.*
For all $v \in H_\mathcal{M}(\Omega)$ and for all $\sigma \in \mathcal{E}_{int}$ with $\mathcal{M}_\sigma = \{K, L\}$ we define $D_\sigma v = |v_K - v_L|$ and $d_\sigma = d_{K,\sigma} + d_{L,\sigma}$, and for all $\sigma \in \mathcal{E}_{ext}$ with $\mathcal{M}_\sigma = \{K\}$, we set $D_\sigma v = |v_K|$ and $d_\sigma = d_{K,\sigma}$. We then define the following family of norms

$$\|v\|_{1,p,\mathcal{M}}^p = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} m(\sigma) d_{K,\sigma} (\frac{D_\sigma v}{d_\sigma})^p; \tag{27}$$

so that in particular

$$\|v\|_{1,2,\mathcal{M}}^2 = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} m(\sigma) d_{K,\sigma} (\frac{D_\sigma v}{d_\sigma})^2.$$

Next we recall two results from [14] which we will use below. The following lemma shows the equivalence between the semi-norm in $X_\mathcal{D}$ and the L^2-norm of the discrete gradient.

Lemma 2.2. *Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.2, and let $\theta \geqslant \theta_\mathcal{D}$ be given. Then there exists $C_1 > 0$ and $C_2 > 0$ only depending on θ and d such that*

$$C_1 |v|_X \leqslant \|\nabla_\mathcal{D} v\|_{L^2(\Omega)} \leqslant C_2 |v|_X \quad \text{for all } v \in X_\mathcal{D}.$$

Lemma 2.3. *Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.2, then there holds*

$$\|v\|_{1,2,\mathcal{M}} \leqslant |v|_X \quad \text{for all } v \in X_{\mathcal{D},0}.$$

Next we show that the bilinear forms defined in (16) and (17) satisfy continuity and coercivity properties.

Lemma 2.4. *Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.2, and let $\theta \geqslant \theta_{\mathcal{D}}$ be given, then:*

(i) There exist positive constants C_1 and α which do not depend on h such that

$$| < u, v >_F | \leqslant C_1 |u|_X |v|_X$$

and

$$< u, u >_F \geqslant \alpha |u|_X^2$$

for all $u, v \in X_{\mathcal{D}}$.

(ii) There exist a positive constant C_2 which does not depend on h that

$$| < u, v >_T | \leqslant C_2 |u|_X |v|_X$$

and

$$< u, u >_T \geqslant 0$$

for all $u, v \in X_{\mathcal{D},0}$.

Proof. (i) Using the definition of the numerical flux (23) and in view of $(\mathcal{H}_2)$ and Lemma 2.2

$$| < u, v >_F | = | \int_\Omega \nabla_{\mathcal{D}} u \cdot \boldsymbol{\Lambda}(x) \nabla_{\mathcal{D}} v| \, d\mathbf{x} \leqslant \overline{\lambda} \|\nabla_{\mathcal{D}} u\|_{L^2(\Omega)} \|\nabla_{\mathcal{D}} v\|_{L^2(\Omega)} \leqslant C_1 |u|_X |v|_X;$$

on the other hand we have that

$$< u, u >_F = \int_\Omega \nabla_{\mathcal{D}} u \cdot \boldsymbol{\Lambda}(x) \nabla_{\mathcal{D}} u \, d\mathbf{x} \geqslant \underline{\lambda} \, d\mathbf{x} \|\nabla_{\mathcal{D}} u\|_{L^2(\Omega)}^2 \geqslant C_2 |u|_X^2.$$

(ii) By the definition (17) we have that

$$< u, v >_T = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} (v_K - v_\sigma) V_{K,\sigma} \overline{u_{K,\sigma}}.$$

Using the definition (8) one can write:

$$< u, v >_T = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K, V_{K,\sigma} \geqslant 0} V_{K,\sigma}(v_K - v_\sigma) u_K + \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K, V_{K,\sigma} \leqslant 0} V_{K,\sigma}(v_K - v_\sigma) u_\sigma,$$

which implies

$$< u, v >_T = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} V_{K,\sigma}(v_K - v_\sigma)u_K - \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K, V_{K,\sigma} \leqslant 0} V_{K,\sigma}(v_K - v_\sigma)(u_K - u_\sigma). \quad (28)$$

Using the Cauchy-Schwarz inequality and the bound $d_{K,\sigma} \leqslant h_{\mathcal{D}}$ we have that

$$| < u, v >_T | \leqslant \sqrt{d} \cdot \|\mathbf{V}\|_{L^\infty(\Omega)} (\sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} m(\sigma)\frac{(v_K - v_\sigma)^2}{d_{K,\sigma}})^{\frac{1}{2}} (\sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} \frac{m(\sigma)d_{K,\sigma}}{d}u_K^2)^{\frac{1}{2}}$$
$$+ h_{\mathcal{D}} \cdot \|\mathbf{V}\|_{L^\infty(\Omega)} (\sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} m(\sigma)\frac{(v_K - v_\sigma)^2}{d_{K,\sigma}})^{\frac{1}{2}} (\sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} m(\sigma)\frac{(u_K - u_\sigma)^2}{d_{K,\sigma}})^{\frac{1}{2}}.$$

and
$$| < u, v >_T | \leqslant \|\mathbf{V}\|_{L^\infty(\Omega)}(\sqrt{d} \cdot |v|_X \|u\|_{L^2(\Omega)} + diam(\Omega) \cdot |v|_X |u|_X)$$

since $\sum_{\sigma \in \mathcal{E}_K} d_{K\sigma}m(\sigma) = m(K)d$. In view of Lemma 2.3 and the discrete Poincaré inequality implied by Lemma 5.1 below we conclude that

$$| < u, v >_T | \leqslant C_2 |u|_X |v|_X.$$

In order to prove the positivity, we write $< u, u >_T$ in the form (28)

$$< u, u >_T = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} V_{K,\sigma}(u_K - u_\sigma)u_K - \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K, V_{K,\sigma} \leqslant 0} V_{K,\sigma}(u_K - u_\sigma)^2;$$

using the algebraic inequality $-2ab \geqslant -a^2 - b^2$ and the discrete boundary condition we obtain

$$< u, u >_T \geqslant \frac{1}{2} \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} V_{K,\sigma}(u_K^2 - u_\sigma^2) = \frac{1}{2} \sum_{K \in \mathcal{M}} u_K^2 \sum_{\sigma \in \mathcal{E}_K} V_{K,\sigma}.$$

By the assumption $(\mathcal{H}_3)$ one has that $\sum_{\sigma \in \mathcal{E}_K} V_{K,\sigma} \geqslant 0$ and we finally conclude that

$$< u, u >_T \geqslant 0.$$

Next we recall a technical lemma presented in [17], Lemma 8.2, which will be useful for the a priori estimates of the next section

Lemma 2.5. Let $B(s)$, $s \in \mathbb{R}$ be defined by

$$B(s) = \beta(s)s - \int_0^s \beta(\tau)d\tau,$$

with β satisfying hypothesis $(\mathcal{H}_1)$. Then $B(s) \geqslant \frac{1}{2}s^2\underline{\beta}$.

3 A priori estimates.

We define below an approximate solution of Problem (1)-(3).

Definition 3.1. (Approximate solution)
Let the sequence of $\{u^n\} \in X_{\mathcal{D},0}^N$, $n \in \{1, \ldots, N\}$, be a solution of the discrete problem (14)-(17), with $\delta t = T/N > 0$. We say that the piecewise constant function $u_{\mathcal{D},\delta t} : \Omega \times [0, T] \to \mathbb{R}$ is an approximate solution of Problem $(\mathcal{P})$ if

$$
\begin{aligned}
u_{\mathcal{D},\delta t}(\mathbf{x}, 0) &= u_K^0 \quad \text{for all } \mathbf{x} \in K, \\
u_{\mathcal{D},\delta t}(\mathbf{x}, t) &= u_K^n \quad \text{for all } (\mathbf{x}, t) \in K \times (t_{n-1}, t_n];
\end{aligned}
$$

we also define its approximate gradient by

$$
\nabla_{\mathcal{D},\delta t} u_{\mathcal{D},\delta t}(\mathbf{x}, t) = \nabla_{\mathcal{D}} u^n(\mathbf{x}) \quad \text{for all } (\mathbf{x}, t) \in K \times (t_{n-1}, t_n].
$$

Lemma 3.1. (A priori estimate) *Let $u_{\mathcal{D},\delta t}$ be an approximate solution of Problem (1)-(3), then it is such that*

$$
\frac{1}{4}\underline{\beta}\|u_{\mathcal{D},\delta t}\|_{L^\infty(0,T;L^2(\Omega))}^2 \leqslant C_1 \quad \text{and} \quad \frac{1}{2}\underline{\lambda}\|\nabla_{\mathcal{D},\delta t} u_{\mathcal{D},\delta t}\|_{L^2(Q_T)}^2 \leqslant C_1, \tag{29}
$$

where

$$
C_1 = \frac{1}{\underline{\beta}}\|\beta(u_0)\|_{L^2(\Omega)}^2 + m(\Omega)TM|\min_{0 \leqslant u \leqslant M} F(u)| + \frac{T}{\underline{\beta}}\|q\|_{L^2(Q_T)}^2;
$$

moreover there exists $C_2 > 0$, such that

$$
\|\beta(u_{\mathcal{D},\delta t})\|_{L^\infty(0,T;L^2(\Omega))} \leqslant C_2. \tag{30}
$$

Proof. Let $m \in [1, N]$ be an arbitrary integer. Summing on $n \in \{1, \ldots, m\}$ the equation (15) with $v^n = u^n$ for each n we obtain

$$
\begin{aligned}
\sum_{K \in \mathcal{M}} m(K) \sum_{n=1}^m u_K^n(\beta(u_K^n) - \beta(u_K^{n-1})) &+ \sum_{n=1}^m \delta t(< u^n, u^n >_F + < u^n, u^n >_T) \\
+ \sum_{n=1}^m \delta t \sum_{K \in \mathcal{M}} m(K) u_K^n F(u_K^n) &= \sum_{n=1}^m \sum_{K \in \mathcal{M}} \delta t \, m(K) u_K^n q_K^n.
\end{aligned}
$$

Next, we consider the function B from Lemma 2.5 defined by

$$
B(u) = \beta(u)u - \int_0^u \beta(\tau)d\tau.
$$

One can see that the following relation holds

$$
B(u_K^n) - B(u_K^{n-1}) = u_K^n(\beta(u_K^n) - \beta(u_K^{n-1})) - \int_{u_K^{n-1}}^{u_K^n}(\beta(\tau) - \beta(u_K^{n-1}))d\tau
$$

and since β is nondecreasing we have that

$$\int_{u_K^{n-1}}^{u_K^n} (\beta(\tau) - \beta(u_K^{n-1}))d\tau \geqslant 0,$$

which implies

$$\begin{aligned}\sum_{K\in\mathcal{M}} m(K)(B(u_K^m) - B(u_K^0)) &= \sum_{K\in\mathcal{M}} m(K)\sum_{n=1}^{m}(B(u_K^n) - B(u_K^{n-1}))\\ &\leqslant \sum_{K\in\mathcal{M}} m(K)\sum_{n=1}^{m} u_K^n(\beta(u_K^n) - \beta(u_K^{n-1})).\end{aligned}$$

In view of Lemma 2.5 we have that

$$\frac{1}{2}\underline{\beta}u^2 \leqslant B(u) \leqslant u\beta(u) \leqslant \frac{(\beta(u))^2}{\underline{\beta}},$$

which yields

$$\frac{1}{2}\underline{\beta}\|u_{\mathcal{D},\delta t}(\cdot,t_m)\|_{L^2(\Omega)}^2 - \frac{1}{\underline{\beta}}\|\beta(u_0)\|_{L^2(\Omega)}^2 \leqslant \sum_{K\in\mathcal{M}} m(K)\sum_{n=1}^{m} u_K^n(\beta(u_K^n) - \beta(u_K^{n-1})).$$

We remark that in view of the hypothesis $(\mathcal{H}_5)$ one has

$$uF(u) \geqslant M\min_{0\leqslant u\leqslant M} F(u),$$

since $\min_{0\leqslant u\leqslant M} F(u) \leqslant 0$. The last statement of Lemma 2.4 implies $< u^n, u^n >_T \geqslant 0$. By the equation (23) and $(\mathcal{H}_2)$ we finally conclude that

$$\begin{aligned}\frac{1}{2}\underline{\beta}\|u_{\mathcal{D},\delta t}(\cdot,t_m)\|_{L^2(\Omega)}^2 &+ \underline{\lambda}\|\nabla_{\mathcal{D},\delta t}u_{\mathcal{D},\delta t}\|_{L^2(\Omega\times(0,km))}^2\\ &\leqslant C + \sum_{n=1}^{m}\sum_{K\in\mathcal{M}} \delta t m(K)u_K^n q_K^n,\end{aligned} \tag{31}$$

where

$$C = \frac{1}{\underline{\beta}}\|\beta(u_0)\|_{L^2(\Omega)}^2 + m(\Omega)TM|\min_{0\leqslant u\leqslant M} F(u)|.$$

Applying Cauchy-Schwarz and Young's inequality to the last term in (31) leads to

$$\begin{aligned}\frac{1}{2}\underline{\beta}\|u_{\mathcal{D},\delta t}(\cdot,t_m)\|_{L^2(\Omega)}^2 + \underline{\lambda}\|\nabla_{\mathcal{D},\delta t}u_{\mathcal{D},\delta t}\|_{L^2(\Omega\times(0,km))}^2 &\leqslant C + \|u_{\mathcal{D},\delta t}\|_{L^2(Q_T)}\|q\|_{L^2(Q_T)}\\ &\leqslant C + \frac{\varepsilon}{2}\|u_{\mathcal{D},\delta t}\|_{L^2(Q_T)}^2 + \frac{1}{2\varepsilon}\|q\|_{L^2(Q_T)}^2.\end{aligned}$$

We then obtain

$$\frac{1}{2}\underline{\beta}\|u_{\mathcal{D},\delta t}\|_{L^\infty(0,T;L^2(\Omega))}^2 \leqslant C + \frac{\varepsilon}{2}T\|u_{\mathcal{D},\delta t}\|_{L^\infty(0,T;L^2(\Omega))}^2 + \frac{1}{2\varepsilon}\|q\|_{L^2(Q_T)}^2$$

13

and
$$\underline{\lambda}\|\nabla_{\mathcal{D},\delta t}u_{\mathcal{D},\delta t}\|^2_{L^2(Q_T)} \leqslant C + \frac{\varepsilon}{2}T\|u_{\mathcal{D},\delta t}\|^2_{L^\infty(0,T;L^2(\Omega))} + \frac{1}{2\varepsilon}\|q\|^2_{L^2(Q_T)}.$$

We now choose $\varepsilon = \underline{\beta}/(2T)$, which gives

$$\frac{1}{4}\underline{\beta}\|u_{\mathcal{D},\delta t}\|^2_{L^\infty(0,T;L^2(\Omega))} \leqslant C \text{ and } \frac{\lambda}{2}\|\nabla_{\mathcal{D},\delta t}u_{\mathcal{D},\delta t}\|^2_{L^2(Q_T)} \leqslant C,$$

where

$$C = \frac{1}{\underline{\beta}}\|\beta(u_0)\|^2_{L^2(\Omega)} + m(\Omega)TM|\min_{0\leqslant u\leqslant M}F(u)| + \frac{T}{\underline{\beta}}\|q\|^2_{L^2(Q_T)}.$$

In order to prove the estimate on $\|\beta(u_{\mathcal{D},\delta t})\|_{L^\infty(0,T;L^2(\Omega))}$ we split β into a bounded and a Lipschitz continuous part by setting $\beta = \beta_1 + \beta_2$, where

$$\beta_1(s) = \begin{cases} \beta(s) & |s| \leqslant P \\ 0 & |s| > P, \end{cases} \qquad \beta_2(s) = \begin{cases} 0 & |s| \leqslant P \\ \beta(s) & |s| > P, \end{cases} \tag{32}$$

and

$$y(s) = \begin{cases} \dfrac{\beta(P) - \beta(-P)}{2P}s + \dfrac{\beta(P) + \beta(-P)}{2} & |s| \leqslant P \\ 0 & |s| > P. \end{cases} \tag{33}$$

We finally define

$$\widetilde{\beta}_1 = \beta_1 - y \text{ and } \widetilde{\beta}_2 = \beta_2 + y. \tag{34}$$

we then remark that $\beta = \widetilde{\beta}_1 + \widetilde{\beta}_2$; we remark that $\widetilde{\beta}_1$ and $\widetilde{\beta}_2$ are continuous and that $\widetilde{\beta}_1$ is bounded by $2C_\beta$, while $\widetilde{\beta}_2$ is Lipschitz continuous with Lipschitz constant $L_{\widetilde{\beta}} = \max(\overline{\beta}, (\beta(P) - \beta(-P))/2P)$. Which implies the $L^\infty(0,T;L^2(\Omega))$ estimate

$$\|\beta(u_{\mathcal{D},\delta t})\|_{L^\infty(0,T;L^2(\Omega))} \leqslant \|\widetilde{\beta}_1(u_{\mathcal{D},\delta t})\|_{L^\infty(0,T;L^2(\Omega))}$$

$$+\|\widetilde{\beta}_2(u_{\mathcal{D},\delta t}) - \widetilde{\beta}_2(0)\|_{L^\infty(0,T;L^2(\Omega))} + \|\widetilde{\beta}_2(0)\|_{L^\infty(0,T;L^2(\Omega))},$$

so that

$$\|\beta(u_{\mathcal{D},\delta t})\|_{L^\infty(0,T;L^2(\Omega))} \leqslant 2m(\Omega)^{\frac{1}{2}}C_\beta + L_{\widetilde{\beta}}\|u_{\mathcal{D},\delta t}\|_{L^\infty(0,T;L^2(\Omega))} + m(\Omega)^{\frac{1}{2}}|\widetilde{\beta}_2(0)|.$$

Remark 3.1. (**Extended discrete problem**) *Let $s > 0$ and $w \in H_{\mathcal{M}}(\Omega)$ (sf. Definition 2.5), we consider the following extended one step problem. Find $u \in X_{\mathcal{D},0}$ such that for all $v \in X_{\mathcal{D},0}$:*

$$s\sum_{K\in\mathcal{M}}m(K)v_K\frac{\beta(u_K) - \beta(w_K)}{\delta t} + <v,u>_F +s<v,u>_T$$

$$+s\sum_{K\in\mathcal{M}}m(K)v_KF(u_K) = s\sum_{K\in\mathcal{M}}m(K)v_Kq_K. \tag{35}$$

14

It can be shown that the solution of the extended problem (35) is bounded in the norm $|\cdot|_X$. More precisely, it is such that

$$\frac{1}{2}\underline{\lambda}\|\nabla_{\mathcal{D},\delta t}u_{\mathcal{D},\delta t}\|_{L^2(Q_T)}^2 \leqslant s\left(\frac{1}{\underline{\beta}}\|\beta(w)\|_{L^2(\Omega)}^2 + m(\Omega)TM|\min_{0\leqslant u\leqslant M}F(u)| + \frac{T}{\underline{\beta}}\|q\|_{L^2(Q_T)}^2\right). \quad (36)$$

Theorem 3.1. (Existence of a discrete solution) *The problem (14)-(17) has at least one solution.*

Proof. Let $(e_i)_{1\leqslant i\leqslant card(X_{\mathcal{D},0})}$ be a family elements of $X_{\mathcal{D},0}$, which components are defined by $(e_i)_j = \delta_{ij}$, where δ_{ij} is the Kronecker symbol. The system of nonlinear equations (6)-(12) may be written in the form

$$E(\beta(u^n) - \beta(u^{n-1})) + \mathcal{A}u^n + \mathcal{C}u^n + \delta t E F(u^n) = \mathcal{Q}^n, \quad (37)$$

where

(i) u^n, $u^{n-1} \in X_{\mathcal{D},0}$;

(ii) E is the diagonal matrix of the size $card(\mathcal{M}) + card(\mathcal{E}_{int})$ with elements

$$(E)_{K,K} = m(K) \quad \text{and} \quad (E)_{\sigma,\sigma} = 0$$

for all $K \in \mathcal{M}, \sigma \in \mathcal{E}_{int}$;

(iii) β and F are continuous mappings from $X_{\mathcal{D},0}$ to itself naturally defined by

$$(\beta(u))_i = \beta(u_i) \quad \text{and} \quad (F(u))_i = F(u_i);$$

(iv) $\mathcal{A}$ and $\mathcal{C}$ are the diffusion matrix and the convection matrix respectively, with components

$$\mathcal{A}_{ij} = \delta t < e_i, e_j >_F \quad \text{and} \quad \mathcal{C}_{ij} = \delta t < e_i, e_j >_T,$$

(v) $\mathcal{Q}^n \in X_{D,0}$ is the source term, given by

$$\mathcal{Q}_K^n = \delta t m(K)q_K^n \quad \text{and} \quad \mathcal{Q}_\sigma^n = 0$$

for all $K \in \mathcal{M}, \sigma \in \mathcal{E}_{int}$;

Due to the coercivity of the bilinear form corresponding to the diffusion the matrix $\mathcal{A}$ is invertible; hence (37) is equivalent to

$$u^n + \mathcal{A}^{-1}(E(\beta(u^n) - \beta(u^{n-1})) + \mathcal{C}u^n + \delta t E F(u^n) - \mathcal{Q}^n) = 0.$$

As it has been done in (35), we introduce the extended formulation

$$u^n + s\mathcal{A}^{-1}(E(\beta(u^n) - \beta(u^{n-1})) + \mathcal{C}u^n + \delta t E F(u^n) - \mathcal{Q}^n) = 0, \quad (38)$$

with $s \in [0, 1]$. Moreover for a given u^{n-1} we define a continuous mapping $H_n : [0, 1] \times X_{\mathcal{D},0} \to X_{\mathcal{D},0}$ by

$$H_n(s, u) = s\mathcal{A}^{-1}(E(\beta(u) - \beta(u^{n-1})) + \mathcal{C}u + \delta t EF(u) - \mathcal{Q}^n).$$

Then the equation (38) can be written in the form $u^n + H_n(s, u^n) = 0$. In view of Remark 3.1, the estimate (30) and the Lemma 2.2 we have that

$$\delta t |u|_X^2 \leqslant C,$$

with some positive constant C, which does not depend on s. Setting $R = \sqrt{C/\delta t + 1}$ we deduce that

$$|u|_X < R \text{ for all } (s, u) \in [0, 1] \times X_{\mathcal{D},0} \text{ such that } u + H_n(s, u) = 0.$$

Therefore the equation $u + H_n(s, u) = 0$ has no solutions on the boundary of the ball B_R of radius R for $s \in [0, 1]$. Next, we denote by $d(Id + H_n(s, \cdot), B_R, 0)$ the topological degree of the application $Id + H_n(s, \cdot)$ with respect to the ball B_R and right-hand side 0. In view of the homotopy invariance of the topological degree and thanks to the fact that $H_n(0, u) = 0$ for all $u \in X_{\mathcal{D},0}$ we have that

$$d(Id + H_n(s, \cdot), B_R, 0) = d(Id + H_n(0, \cdot), B_R, 0) = 1 \text{ for all } s \in [0, 1],$$

where we have applied [[11], Theorem 3.1 (d1) and (d3)]. Thus, by [[11], Theorem 3.1 (d4)], there exists u^n such that $u^n + H_n(1, u^n) = 0$, so that u^n is a solution of (37).

Theorem 3.2. (Uniqueness of the discrete solution) *The solution of the problem (14)-(17) is unique.*

Proof. We give a proof by contradiction. Let $u_{\mathcal{D},\delta t}$ and $\widetilde{u}_{h,k}$ be two different solutions of (14)-(17), such that $u^m = \widetilde{u}^m$ for all $m = 1, \dots, n - 1$, but $u^n \neq \widetilde{u}^n$. We define $r^n = u^n - \widetilde{u}^n$. In view of (15) with $v = r^n$ we have that

$$\sum_{K \in \mathcal{M}} m(K) r_K^n \frac{\beta(u_K^n) - \beta(\widetilde{u}_K^n)}{\delta t} + <r^n, r^n>_F + <r^n, r^n>_T$$

$$+ \sum_{K \in \mathcal{M}} m(K) r_K^n (F(u_K^n) - F(\widetilde{u}_K^n)) = 0.$$

We apply Lemma 2.4 as well as the assumptions $(\mathcal{H}_1)$ and $(\mathcal{H}_5)$ in order to estimate each term in the above equation. We obtain that

$$(\underline{\beta}/\delta t - \overline{F}) \sum_{K \in \mathcal{M}} m(K)(r_K^n)^2 + \alpha |r^n|_X^2 \leqslant 0,$$

where α is the coercivity constant. Finally, in view of the assumption (5) on the time step we deduce that

$$|r^n|_X = 0.$$

4 Estimate on time translates

To begin with we give two technical lemmas which will be useful for proving the estimate on time translates

Lemma 4.1. *Let $T > 0$, $\tau \in (0, T)$, $N \in \mathbb{N} \setminus \{0\}$, $\delta t = T/N$ be given and $(a^n)_{n \in \mathbb{N} \setminus \{0\}}$ be a family of non negative real values. Let $\lceil s \rceil$ denotes the smallest integer larger or equal to s. Then*

$$\int_0^{T-\tau} \sum_{\lceil t/\delta t \rceil + 1 \leqslant n \leqslant \lceil (t+\tau)/\delta t \rceil} a^n \, dt \leqslant \tau \sum_{n=1}^N a^n.$$

Proof. One has that

$$\int_0^{T-\tau} \sum_{\lceil t/\delta t \rceil + 1 \leqslant n \leqslant \lceil (t+\tau)/\delta t \rceil} a^n \, dt \leqslant \int_0^{T-\tau} \sum_{t/\delta t + 1 \leqslant n < (t+\tau)/\delta t + 1} a^n \, dt = \int_0^{T-\tau} \sum_{t \leqslant m\delta t < t+\tau} a^{m+1} \, dt$$

We remark that if $\lceil t/\delta t \rceil + 1 > \lceil (t+\tau)/\delta t \rceil$, then the above inequality seal holds, with the left hand side term equal to zero. We define a characteristic function $\chi(n, t_1, t_2)$ by

$$\chi(n, t_1, t_2) = \begin{cases} 1 & \text{if } t_1 \leqslant n\delta t < t_2, \\ 0 & \text{otherwise.} \end{cases}$$

Then we obtain that

$$\int_0^{T-\tau} \sum_{\lceil t/\delta t \rceil + 1 \leqslant n \leqslant \lceil (t+\tau)/\delta t \rceil} a^n \, dt \leqslant \sum_{m=1}^{N-1} a^{m+1} \int_0^{T-\tau} \chi(n, t, t+\tau) \, dt \leqslant \tau \sum_{m=1}^N a^m.$$

Lemma 4.2. *Let $T > 0$, $\tau \in (0, T)$, $N \in \mathbb{N} \setminus \{0\}$, $\delta t = T/N$, $\zeta \in [0, \tau]$ be given and $(a^n)_{n \in \mathbb{N} \setminus \{0\}}$ be a family of nonnegative real values. Let $\lceil s \rceil$ denotes the smallest integer larger or equal to s. Then*

$$\int_0^{T-\tau} \sum_{\lceil t/\delta t \rceil + 1 \leqslant n \leqslant \lceil (t+\tau)/\delta t \rceil} a^{\lceil (t+\zeta)/\delta t \rceil} \, dt \leqslant \tau \sum_{n=1}^N a^n.$$

Proof. As in the proof of the previous Lemma we have that

$$\int_0^{T-\tau} \sum_{\lceil t/\delta t \rceil + 1 \leqslant n \leqslant \lceil (t+\tau)/\delta t \rceil} a^{\lceil (t+\zeta)/\delta t \rceil} \, dt \leqslant \sum_{n=1}^N \int_0^{T-\tau} a^{\lceil (t+\zeta)/\delta t \rceil} \chi(n, t, t+\tau) \, dt$$

A simple change of variable implies

$$\begin{aligned}
\sum_{n=1}^N \int_0^{T-\tau} a^{\lceil (t+\zeta)/\delta t \rceil} \chi(n, t, t+\tau) \, dt &\leqslant \sum_{n=1}^N \int_0^T a^{\lceil s/\delta t \rceil} \chi(n, s - \zeta, s - \zeta + \tau) \, ds \\
&= \sum_{m=1}^N a^m \sum_{n=1}^N \int_{(m-1)\delta t}^{m\delta t} \chi(n, s - \zeta, s - \zeta + \tau) \, ds.
\end{aligned}$$

In order to conclude the proof we remark that $\chi(n, t_1, t_2) = \chi(n + m, t_1 + m\delta t, t_2 + m\delta t)$ for all $n, m \in \mathbb{Z}$, $t_1, t_2 \in \mathbb{R}$, which in turn implies that

$$
\sum_{n=1}^{N} \int_{(m-1)\delta t}^{m\delta t} \chi(n, s - \zeta, s - \zeta + \tau) ds = \sum_{n=1}^{N} \int_{-(n+1)\delta t}^{-n\delta t} \chi(-m, s - \zeta, s - \zeta + \tau) ds
$$

$$
\leqslant \int_{\mathbb{R}} \chi(-m, s - \zeta, s - \zeta + \tau) ds = \tau.
$$

Theorem 4.1. *Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.2 and let $\{u_{\mathcal{D},\delta t}\}$ be a solution of the discrete problem in sense of Definition 3.1. Let also $\theta \geqslant \theta_{\mathcal{D}}$ be given. Then there exists a positive constant C only depending on θ such that*

$$
\int_0^{T-\tau} \int_\Omega (u_{\mathcal{D},\delta t}(x, t+\tau) - u_{\mathcal{D},\delta t}(x, t))^2 dx dt \leqslant C\tau, \tag{39}
$$

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

Proof. To begin with we use the hypothesis $(\mathcal{H}_1)$ to obtain

$$
\underline{\beta} \int_0^{T-\tau} \int_\Omega (u_{\mathcal{D},\delta t}(x, t+\tau) - u_{\mathcal{D},\delta t}(x, t))^2 \, d\mathbf{x} dt
$$

$$
= \underline{\beta} \int_0^{T-\tau} \sum_{K \in \mathcal{M}} m(K) (u_K^{\lceil (t+\tau)/\delta t \rceil} - u_K^{\lceil t/\delta t \rceil})^2 dt
$$

$$
\leqslant \int_0^{T-\tau} \sum_{K \in \mathcal{M}} m(K) (u_K^{\lceil (t+\tau)/\delta t \rceil} - u_K^{\lceil t/\delta t \rceil})(\beta(u_K^{\lceil (t+\tau)/\delta t \rceil}) - \beta(u_K^{\lceil t/\delta t \rceil})) dt
$$

$$
= \int_0^{T-\tau} \sum_{K \in \mathcal{M}} m(K) (u_K^{\lceil (t+\tau)/\delta t \rceil} - u_K^{\lceil t/\delta t \rceil}) \sum_{\lceil t/\delta t \rceil + 1 \leqslant n \leqslant \lceil (t+\tau)/\delta t \rceil} m(K)(\beta(u_K^n) - \beta(u_K^{n-1})) dt
$$

For a given k and for all real t and τ we define the following set

$$
n(t, \tau) = \{n \in \mathbb{N}, \ \lceil t/\delta t \rceil + 1 \leqslant n \leqslant \lceil (t+\tau)/\delta t \rceil\},
$$

which can be empty. Then, the discrete equation (7) implies

$$
\underline{\beta} \int_0^{T-\tau} \int_\Omega (u_{\mathcal{D},\delta t}(x, t+\tau) - u_{\mathcal{D},\delta t}(x, t))^2 \, d\mathbf{x} dt \leqslant \int_0^{T-\tau} \sum_{K \in \mathcal{M}} (u_K^{\lceil (t+\tau)/\delta t \rceil} - u_K^{\lceil t/\delta t \rceil})
$$

$$
\cdot \sum_{n \in n(t,\tau)} \delta t (m(K) q_K^n - \sum_{\sigma \in \mathcal{E}_K} (F_{K,\sigma}(u^n) + V_{K,\sigma} \overline{u_{K,\sigma}^n}) - m(K) F(u_K^n)) dt.
$$

Let us define the expressions $A_{D,C}$, A_R and A_S by

$$A_{D,C} = \int_0^{T-\tau} \sum_{n \in n(t,\tau)} \delta t \sum_{K \in \mathcal{M}} (u_K^{\lceil (t+\tau)/\delta t \rceil} - u_K^{\lceil t/\delta t \rceil}) \sum_{\sigma \in \mathcal{E}_K} (F_{K,\sigma}(u^n) + V_{K,\sigma} \overline{u_{K,\sigma}^n}) dt,$$

$$A_R = \int_0^{T-\tau} \sum_{n \in n(t,\tau)} \delta t \sum_{K \in \mathcal{M}} m(K)(u_K^{\lceil (t+\tau)/\delta t \rceil} - u_K^{\lceil t/\delta t \rceil}) F(u_K^n) dt,$$

$$A_S = \int_0^{T-\tau} \sum_{n \in n(t,\tau)} \delta t \sum_{K \in \mathcal{M}} m(K)(u_K^{\lceil (t+\tau)/\delta t \rceil} - u_K^{\lceil t/\delta t \rceil}) q_K^n dt,$$

which we will estimate below. In view of (13), (16) and (17) we obtain

$$A_{D,C} = \int_0^{T-\tau} \sum_{n \in n(t,\tau)} \delta t(< u^{\lceil (t+\tau)/\delta t \rceil} - u^{\lceil t/\delta t \rceil}, u^n >_F + < u^{\lceil (t+\tau)/\delta t \rceil} - u^{\lceil t/\delta t \rceil}, u^n >_T) dt$$

In view of Lemma 2.4 we have that $| < u, v >_F | + | < u, v >_T | \leqslant C|u|_X |v|_X$ for all $u, v \in X_{\mathcal{D},0}$ and since $2ab \leqslant a^2 + b^2$ one has

$$|A_{D,C}| \leqslant C \int_0^{T-\tau} \sum_{n \in n(t,\tau)} \delta t(|u^{\lceil (t+\tau)/\delta t \rceil}|_X + |u^{\lceil t/\delta t \rceil}|_X)|u^n|_X dt$$

$$\leqslant C(\int_0^{T-\tau} \sum_{n \in n(t,\tau)} \delta t|u^{\lceil t/\delta t \rceil}|_X^2 + \int_0^{T-\tau} \sum_{n \in n(t,\tau)} \delta t|u^{\lceil (t+\tau)/\delta t \rceil}|_X^2 + \int_0^{T-\tau} \sum_{n \in n(t,\tau)} \delta t|u^n|_X^2 dt).$$

It follows from the estimate (29) and the Lemmas 4.1 and 4.2 that

$$|A_{D,C}| \leqslant \tau C \sum_{n=1}^N \delta t|u^n|_X^2 \leqslant C\tau.$$

Next, we estimate the term A_R; we remark that for all $u, v \in \mathbb{R}$ it holds

$$vF(u) \leqslant L_F|v||u| \leqslant \frac{1}{2}v^2 + \frac{1}{2}L_F^2 u^2 \quad \text{if } u < 0,$$

$$vF(u) \leqslant |v| \max_{0 \leqslant u \leqslant M} |F(u)| \leqslant \frac{1}{2}v^2 + \frac{1}{2} \max_{0 \leqslant u \leqslant M} F^2(u) \quad \text{if } 0 \leqslant u \leqslant M,$$

$$vF(u) \leqslant |v|(|F(u) - F(M)| + |F(M)|) \leqslant |v|(L_F|u - M| + F(M))$$

$$\leqslant v^2 + \frac{1}{2}L_F^2|u|^2 + \frac{1}{2}(L_F M + F(M))^2 \quad \text{if } u > M.$$

Hence,

$$\sum_{K \in \mathcal{M}} m(K)v_K F(u_K) \leqslant \|v\|_{L^2(\Omega)}^2 + \frac{1}{2}L_F^2\|u\|_{L^2(\Omega)}^2 + \frac{1}{2}C_F$$

for all $v, u \in H_{\mathcal{M}}$, where $C_F = \frac{1}{2}m(\Omega)(\max_{0 \leqslant u \leqslant M} F^2(u) + (L_F M + F(M))^2)$. We obtain

$$
\begin{aligned}
|A_R| \quad \leqslant \quad & \int_0^{T-\tau} \sum_{n \in n(t,\tau)} \delta t(\|u_{\mathcal{D},\delta t}(\cdot, \lceil t/\delta t \rceil)\|_{L^2(\Omega)}^2 + \|u_{\mathcal{D},\delta t}(\cdot, \lceil (t+\tau)/\delta t \rceil)\|_{L^2(\Omega)}^2)dt \\
& + \int_0^{T-\tau} \sum_{n \in n(t,\tau)} \delta t(L_F^2 \|u_{\mathcal{D},\delta t}(\cdot, t_n)\|_{L^2(\Omega)}^2 + C_F)dt.
\end{aligned}
$$

One more time it follows from the estimate (29) and the Lemmas 4.1 and 4.2 that

$$
|A_R| \leqslant \tau \sum_{n=1}^{N} \delta t(C\|u_{\mathcal{D},\delta t}(\cdot, t_n)\|_{L^2(\Omega)}^2 + C_F) \leqslant \tau C
$$

In the same way we proceed for the term $|A_S|$, one has that

$$
A_S = \int_0^{T-\tau} \sum_{n \in n(t,\tau)} \sum_{K \in \mathcal{M}} \int_{t_{n-1}}^{t_n} \int_K (u_K^{\lceil (t+\tau)/\delta t \rceil} - u_K^{\lceil t/\delta t \rceil})q(\mathbf{x}, s) \; d\mathbf{x}dsdt
$$

and

$$
\begin{aligned}
|A_S| \quad \leqslant \quad & \int_0^{T-\tau} \sum_{n \in n(t,\tau)} \sum_{K \in \mathcal{M}} \int_{t_{n-1}}^{t_n} \int_K \frac{1}{2}((u_K^{\lceil t/\delta t \rceil})^2 + (u_K^{\lceil (t+\tau)/\delta t \rceil})^2) + (q(\mathbf{x}, s))^2 \; d\mathbf{x}dsdt \\
= \quad & \int_0^{T-\tau} \sum_{n \in n(t,\tau)} \frac{1}{2}\delta t(\|u_{\mathcal{D},\delta t}(\cdot, \lceil t/\delta t \rceil)\|_{L^2(\Omega)}^2 + \|u_{\mathcal{D},\delta t}(\cdot, \lceil (t+\tau)/\delta t \rceil)\|_{L^2(\Omega)}^2)dt \\
& + \int_0^{T-\tau} \sum_{n \in n(t,\tau)} \sum_{K \in \mathcal{M}} \int_{t_{n-1}}^{t_n} \int_K (q(\mathbf{x}, s))^2 \; d\mathbf{x}dsdt.
\end{aligned}
$$

In view of Lemmas 4.2 and 4.1 we obtain

$$
|A_S| \leqslant \tau(\|u_{\mathcal{D},\delta t}\|_{L^2(Q_T)}^2 + \|q\|_{L^2(Q_T)}^2).
$$

Finally we use an a priori estimate (29) and the hypothesis $(\mathcal{H}_6)$ to conclude the proof.

5 Estimate on space translates

In this section we prove an estimate on the L^2-norm of differences of space translates of the discrete solution. We state without proof two results from [14], which are useful in our study.

Lemma 5.1. *Let $d \geqslant 1$, $1 \leqslant p < \infty$ and Ω be an open bounded connected subset of $\mathbb{R}^d$. Let $\mathcal{D}$ be a mesh of Ω in the sense of Definition 2.2. Let $\eta > 0$ be such that $\eta \leqslant d_{K,\sigma}/d_{L,\sigma} \leqslant 1/\eta$*

for all $\sigma \in \mathcal{M}_\sigma = \{K, L\}$. Then, there exists $q > p$ only depending on p and there exist a positive constant C, only depending on d, Ω, p and η such that:

$$\|u\|_{L^q(\Omega)} \leqslant C\|u\|_{1,p,\mathcal{M}} \tag{40}$$

for all $u \in H_\mathcal{M}(\Omega)$. We recall that $H_\mathcal{M}(\Omega) \subset L^2(\Omega)$ is the set of piecewise constant functions on the control volumes of the mesh.

Lemma 5.2. *Let $d \geqslant 1$ and Ω be a polyhedral open bounded connected subset of $\mathbb{R}^d$. Let $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{P})$ be a discretization of Ω in the sense of Definition 2.2 and let $u \in H_\mathcal{M}(\Omega)$. Then, with notation of Definition 2.5:*

$$\|u(\cdot + \mathbf{y}) - u\|_{L^1(\mathbb{R}^d)} \leqslant |\mathbf{y}|\sqrt{d}\|u\|_{1,1,\mathcal{M}}, \tag{41}$$

where u is defined on the whole $\mathbb{R}^d$, taking $u = 0$ outside Ω.

Next we show that a similar inequality holds in every L^p-norm.

Lemma 5.3. *Let $d \geqslant 1$, $1 \leqslant p < \infty$ and Ω be an open bounded connected subset of $\mathbb{R}^d$ and $T > 0$. Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.2. Let $\eta > 0$ such that $\eta \leqslant d_{K,\sigma}/d_{L,\sigma} \leqslant 1/\eta$ for all $\sigma \in \mathcal{M}_\sigma = \{K, L\}$. There exist $C > 0$ and $\rho > 0$, only depending on d, p, Ω and η such that*

$$\|u(\cdot + \mathbf{y}) - u\|_{L^p(\mathbb{R}^d)} \leqslant C|\mathbf{y}|^\rho\|u\|_{1,p,\mathcal{M}},$$

where u is defined on $\mathbb{R}^d$, taking $u = 0$ outside Ω.

Proof. In view of Lemma 5.1, there exist $q > p$ and a positive constant C such that

$$\|u\|_{L^q(\mathbb{R}^d)} \leqslant C\|u\|_{1,p,\mathcal{M}}. \tag{42}$$

We apply the Interpolation Inequality [[1], Theorem 2.11, p.27]

$$\|u(\cdot + \mathbf{y}) - u\|_{L^p(\mathbb{R}^d)} \leqslant \|u(\cdot + \mathbf{y}) - u\|^\rho_{L^1(\mathbb{R}^d)}\|u(\cdot + \mathbf{y}) - u\|^{1-\rho}_{L^q(\mathbb{R}^d)}, \tag{43}$$

where

$$\rho = \frac{1}{p} \cdot \frac{q - p}{q - 1}.$$

Moreover (42) implies that

$$\|u(\cdot + \mathbf{y}) - u\|_{L^q(\mathbb{R}^d)} \leqslant 2\|u\|_{L^q(\mathbb{R}^d)} \leqslant C\|u\|_{1,p,\mathcal{M}},$$

so that by (41) and (43) implies that

$$\|u(\cdot + \mathbf{y}) - u\|_{L^p(\mathbb{R}^d)} \leqslant C|\mathbf{y}|^\rho(\|u\|_{1,1,\mathcal{M}})^\rho(\|u\|_{1,p,\mathcal{M}})^{1-\rho}.$$

Applying Hölder inequality we obtain that

$$\|u\|_{1,1,\mathcal{M}} \leqslant C\|u\|_{1,p,\mathcal{M}}$$

for some positive constant C. Then

$$\|u(\cdot + \mathbf{y}) - u\|_{L^p(\mathbb{R}^d)} \leqslant C|\mathbf{y}|^\rho\|u\|_{1,p,\mathcal{M}}.$$

Theorem 5.1. *Let $\{\mathcal{D}_h\}$ be a family of discretizations in the sense of Definition 2.2 and let θ be a positive constant such that $\theta_{\mathcal{D}} \leqslant \theta$ for all $\mathcal{D} \in \mathcal{D}_h$. Let $\{u_{\mathcal{D},\delta t}\}$ be a family of approximate solutions corresponding to $\mathcal{D}_h$ and $\delta t = T/N$ for some $N \in \mathbb{N} \setminus \{0\}$. Then $\{u_{\mathcal{D},\delta t}\}$ is relatively compact in $L^2(Q_T)$.*

Proof. To begin with, we extend $u_{\mathcal{D},\delta t}$ by zero outside of Q_T. Applying the Lemma 5.3 with $p = 2$ yields

$$\|u_{\mathcal{D},\delta t}(\cdot + \mathbf{y}, t) - u_{\mathcal{D},\delta t}(\cdot, t)\|_{L^2(\mathbb{R}^d)} \leqslant C|\mathbf{y}|^\rho \|u_{\mathcal{D},\delta t}(\cdot, t)\|_{1,2,\mathcal{M}}$$

for some positive constants $\rho > 0$ and $C > 0$. Integrating on $(0, T)$ we obtain

$$\|u_{\mathcal{D},\delta t}(\cdot + \mathbf{y}, \cdot) - u_{\mathcal{D},\delta t}\|^2_{L^2(\mathbb{R}^d \times (0,T))} \leqslant C|\mathbf{y}|^{2\rho} \sum_{n=1}^N \delta t \|u_{\mathcal{D},\delta t}(\cdot, t_n)\|^2_{1,2,\mathcal{M}}.$$

Then in view of the lemmas 2.3, 2.2 and the estimate (29) we obtain the bound

$$\|u_{\mathcal{D},\delta t}(\cdot + \mathbf{y}) - u_{\mathcal{D},\delta t}\|_{L^2(\mathbb{R}^d \times (0,T))} \leqslant C|\mathbf{y}|^\rho,$$

which, combined with (39) gives

$$\|u_{\mathcal{D},\delta t}(\cdot + \mathbf{y}, \cdot + \tau) - u_{\mathcal{D},\delta t}\|_{L^2(\mathbb{R}^d \times (0,T))}$$

$$\leqslant \|u_{\mathcal{D},\delta t}(\cdot + \mathbf{y}, \cdot + \tau) - u_{\mathcal{D},\delta t}(\cdot + \mathbf{y}, \cdot)\|_{L^2(\mathbb{R}^d \times (0,T))} + \|u_{\mathcal{D},\delta t}(\cdot + \mathbf{y}, \cdot) - u_{\mathcal{D},\delta t}\|_{L^2(\mathbb{R}^d \times (0,T))}$$

$$\leqslant C(\sqrt{\tau} + |\mathbf{y}|^\rho).$$

Then the Fréchet-Kolmogorov Compactness Theorem implies that the family $\{u_{\mathcal{D},\delta t}\}$ is relatively compact in $L^2(\mathbb{R}^d \times (0,T))$ and thus in $L^2(Q_T)$.

6 Convergence

Theorem 6.1. *Let $\{\mathcal{D}_h\}$ be a family of discretizations in the sense of Definition 2.2 and let θ be a positive constant such that $\theta_{\mathcal{D}} \leqslant \theta$ for all $\mathcal{D} \in \mathcal{D}_h$. Let $\{u_{\mathcal{D},\delta t}\}$ be a family of approximate solutions corresponding to $\mathcal{D}_h$ and $\delta t = T/N$ for some $N \in \mathbb{N} \setminus \{0\}$. Then there exists a subsequence of $\{u_{\mathcal{D},\delta t}\}$, which we denote again by $\{u_{\mathcal{D},\delta t}\}$, such that $u_{\mathcal{D},\delta t} \to u$ strongly in $L^2(Q_T)$ as $h_{\mathcal{D}}, \delta t \to 0$, where u is a weak solution of Problem $(\mathcal{P})$. Moreover $u \in L^2(0,T; H_0^1(\Omega))$ and $\nabla_{\mathcal{D},\delta t} u_{\mathcal{D},\delta t}$ weakly converge in $L^2(Q_T)^d$ to ∇u. In the case that F is nondecreasing, the whole sequence $\{u_{\mathcal{D},\delta t}\}$ converges to the unique weak solution u of Problem $(\mathcal{P})$.*

Proof. By Theorem 5.1 there exist a subsequence of $\{u_{\mathcal{D},\delta t}\}$ that we still denote by $\{u_{\mathcal{D},\delta t}\}$ and a function $u \in L^2(Q_T)$ such that $u_{\mathcal{D},\delta t} \to u$ strongly in $L^2(Q_T)$ as $h_{\mathcal{D}}, \delta t \to 0$ (and also in $L^2(\mathbb{R}^d \times (0,T))$ taking $u_{\mathcal{D},\delta t} = 0$ outside of $\Omega \times (0,T)$). In view of (29) there exists

a function $\mathbf{G} \in L^2(Q_T)^d$ such that $\nabla_{\mathcal{D},\delta t} u_{\mathcal{D},\delta t}$ weakly converge in $L^2(Q_T)^d$ to $\mathbf{G}$ along a subsequence as $h_\mathcal{D}, \delta t \to 0$. In order to show that $\mathbf{G} = \nabla u$ we consider an arbitrary vector function $\mathbf{w} \in C([0,T]; C_c^\infty(\mathbb{R}^d))$ and the term $T_\mathbf{G}^1$ defined by

$$T_\mathbf{G}^1 = \int_0^T \int_{\mathbb{R}^d} \nabla_{\mathcal{D},\delta t} u_{\mathcal{D},\delta t}(\mathbf{x},t) \cdot \mathbf{w}(\mathbf{x},t) \; d\mathbf{x}dt.$$

Using the Definition 3.1 and (20) we obtain that $T_G^1 = T_G^2 + T_G^3$, with

$$T_\mathbf{G}^2 = \sum_{n=1}^N \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} \delta t m(\sigma)(u_\sigma^n - u_K^n)\mathbf{n}_{K,\sigma} \cdot \mathbf{w}_K^n$$

and

$$T_\mathbf{G}^3 = \sum_{n=1}^N \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} R_{K,\sigma}(u^n)\mathbf{n}_{K,\sigma} \cdot \int_{t_{n-1}}^{t_n} \int_{D_{K,\sigma}} \mathbf{w}(\mathbf{x},t) \; d\mathbf{x}dt,$$

where $\mathbf{w}_K^n = \dfrac{1}{\delta t m(K)} \int_{t_{n-1}}^{t_n} \int_K \mathbf{w}(\mathbf{x},t) \; d\mathbf{x}dt$. We compare $T_\mathbf{G}^2$ with $T_\mathbf{G}^4$ defined by

$$T_\mathbf{G}^4 = \sum_{n=1}^N \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} \delta t m(\sigma)(u_\sigma^n - u_K^n)\mathbf{n}_{K,\sigma} \cdot \mathbf{w}_\sigma^n,$$

where $\mathbf{w}_\sigma^n = \dfrac{1}{\delta t m(\sigma)} \int_{t_{n-1}}^{t_n} \int_\sigma \mathbf{w}(\mathbf{x},t) \; d\gamma dt$. One can see that

$$(T_\mathbf{G}^2 - T_\mathbf{G}^4)^2 \leqslant \sum_{n=1}^N \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} \frac{\delta t m(\sigma)}{d_{K,\sigma}}(u_\sigma^n - u_K^n)^2 \sum_{n=1}^N \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} \delta t m(\sigma)d_{K,\sigma}|\mathbf{w}_K^n - \mathbf{w}_\sigma^n|^2,$$

which leads to $T_\mathbf{G}^2 \to T_\mathbf{G}^4$ as $h_\mathcal{D} \to 0$. Then,

$$T_\mathbf{G}^4 = -\sum_{n=1}^N \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} \delta t m(\sigma)u_K^n \mathbf{n}_{K,\sigma} \cdot \mathbf{w}_\sigma^n = -\int_0^T \int_{\mathbb{R}^d} u_{\mathcal{D},\delta t}(\mathbf{x},t) \; \nabla \cdot \mathbf{w}(\mathbf{x},t) \; d\mathbf{x}dt.$$

and we conclude that

$$\lim_{h_\mathcal{D},\delta t \to 0} T_\mathbf{G}^4 = -\int_0^T \int_{\mathbb{R}^d} u(x,t)\nabla \cdot \; \mathbf{w}(\mathbf{x},t) \; d\mathbf{x}dt.$$

Next, we show that $\lim\limits_{h_\mathcal{D},\delta t \to 0} T_\mathbf{G}^3 = 0$. Thanks to (22) we have that

$$T_\mathbf{G}^3 = \sum_{n=1}^N \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} R_{K,\sigma}(u^n)\mathbf{n}_{K,\sigma} \cdot \int_{t_{n-1}}^{t_n} \int_{D_{K,\sigma}} (\mathbf{w}(\mathbf{x},t) - \mathbf{w}_K^n) \; d\mathbf{x}dt.$$

Since $\mathbf{w}$ is a regular function, there exist a positive constant $C = C(\mathbf{w})$ such that

$$\left| \int_{t_{n-1}}^{t_n} \int_{D_{K,\sigma}} (\mathbf{w}(\mathbf{x}, t) - \mathbf{w}_K^n) \, d\mathbf{x} dt \right| \leqslant C \delta t \frac{m(\sigma) d_{K,\sigma}}{d} (h_{\mathcal{D}} + \delta t).$$

On the other hand by (21) and in view of regularity of the mesh, we have that

$$(R_{K,\sigma} u^n)^2 \leqslant 2d \left(\left(\frac{u_\sigma^n - u_K^n}{d_{K,\sigma}} \right)^2 + |\nabla_K u^n|^2 \left| \frac{\mathbf{x}_\sigma - \mathbf{x}_K}{d_{K,\sigma}} \right|^2 \right) \leqslant 2d \left(\left(\frac{u_\sigma^n - u_K^n}{d_{K,\sigma}} \right)^2 + \theta^2 |\nabla_K u^n|^2 \right).$$

Applying the Cauchy-Schwarz inequality inequality, we get

$$\lim_{h_{\mathcal{D}}, \delta t \to 0} T_{\mathbf{G}}^3 = 0$$

This implies that the function $\mathbf{G} \in L^2(\mathbb{R}^d \times (0, T))^d$ is a.e. equal to ∇u in $\mathbb{R}^d \times (0, T)$. Since $u = 0$ outside of Ω, it follows that $u \in L^2(0, T; H_0^1(\Omega))$.

Next we show that u is a weak solution of the problem $(\mathcal{P})$. For this purpose, we introduce the function space

$$\Phi = \{\varphi \in C^{2,1}(\overline{\Omega} \times [0, T]), \quad \varphi = 0 \text{ on } \partial\Omega \times [0, T], \quad \varphi(\cdot, T) = 0\}.$$

Taking an arbitrary $\varphi \in \Phi$, we define the sequence of elements of $X_{\mathcal{D},0}$

$$\varphi^n = P_{\mathcal{D}} \varphi(\cdot, t_n) \text{ for all } n \in \{1, \ldots, N\}$$

which implies $\varphi_K^n = \varphi(\mathbf{x}_K, t_n)$ and $\varphi_\sigma^n = \varphi(\mathbf{x}_\sigma, t_n)$. Next setting

$$v^n = \varphi^{n-1} \text{ for all } n \in \{1, \ldots, N\}$$

in (15), we obtain, also in view of (16) and (17) that

$$T_T + T_D + T_C + T_R = T_S,$$

where

$$T_T = \sum_{n=1}^{N} \sum_{K \in \mathcal{M}} m(K)(\beta(u_K^n) - \beta(u_K^{n-1}))\varphi_K^{n-1},$$

$$T_D = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} (\varphi_K^{n-1} - \varphi_\sigma^{n-1})F_{K,\sigma}(u^n),$$

$$T_C = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} (\varphi_K^{n-1} - \varphi_\sigma^{n-1})V_{K,\sigma}\overline{u_{K,\sigma}^n},$$

$$T_R = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} m(K)\varphi_K^{n-1}F(u_K^n)$$

and

$$T_S = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} m(K)\varphi_K^{n-1}q_K^n.$$

We successively search for the limit of each of these terms as $h_{\mathcal{D}}$ and k tend to zero.

6.1 Time evolution term

Using discrete integration by parts and the fact that $\varphi(\mathbf{x}, T) = 0$ we obtain

$$T_T = -\sum_{n=1}^{N} \sum_{K \in \mathcal{M}} m(K)(\varphi_K^n - \varphi_K^{n-1})\beta(u_K^n) - \sum_{K \in \mathcal{M}} m(K)\varphi_K^0 \beta(u_K^0).$$

First we show that

$$\sum_{K \in \mathcal{M}} m(K)\beta(u_K^0)\varphi_K^0 \to \int_{\Omega} \beta(u_0(\mathbf{x}))\varphi(\mathbf{x}, 0) \, d\mathbf{x}.$$

For this purpose we define

$$T_T^0 = \sum_{K \in \mathcal{M}} m(K)\beta(u_K^0)\varphi_K^0 - \int_{\Omega} \beta(u_0(\mathbf{x}))\varphi(\mathbf{x}, 0) \, d\mathbf{x}.$$

Next we subtract $\int_{\Omega} \beta(u_K^0)\varphi(\mathbf{x}, 0)$ from each term to deduce that,

$$T_T^0 = \sum_{K \in \mathcal{M}} \int_K \beta(u_K^0)(\varphi_K^0 - \varphi(\mathbf{x}, 0)) \, d\mathbf{x} - \sum_{K \in \mathcal{M}} \int_K (\beta(u_0(\mathbf{x})) - \beta(u_K^0))\varphi(\mathbf{x}, 0) \, d\mathbf{x}. \quad (44)$$

In view of the regularity of the test function $\varphi \in C^{2,1}(\overline{\Omega} \times [0, T])$ we have that

$$|\varphi_K^0 - \varphi(\mathbf{x}, 0)| \leqslant C h_{\mathcal{D}} \text{ for all } \mathbf{x} \in K$$

and

$$|\varphi(\mathbf{x}, 0)| \leqslant C.$$

We also remark that by (6) one has that $|u_K^0| \leqslant \|u_0\|_{L^\infty(\Omega)}$ and moreover the monotonicity hypothesis $(\mathcal{H}_1)$ implies that $|\beta(u_K^0)| \leqslant \beta(\|u_0\|_{L^\infty(\Omega)})$ for all $K \in \mathcal{M}$; consequently the first term on the right-hand side of (44) tends to zero as $h_{\mathcal{D}} \to 0$ and the second term can be estimated by

$$C \sum_{K \in \mathcal{M}} \int_K |\beta(u_0(\mathbf{x})) - \beta(u_K^0)| \, d\mathbf{x} = C \int_{\Omega} |\beta(u_0(\mathbf{x})) - \beta(u_{\mathcal{D}, \delta t}(\mathbf{x}, 0))| \, d\mathbf{x}.$$

By the discrete initial condition (6) and the Definition 3.1 one has

$$\int_{\Omega} |u_0(\mathbf{x}) - u_{\mathcal{D}, \delta t}(\mathbf{x}, 0)| \, d\mathbf{x} \to 0 \text{ as } h_{\mathcal{D}} \to 0,$$

or in other words $u_{\mathcal{D}, \delta t}(0)$ converges strongly to u_0 in $L^1(\Omega)$ as $h_{\mathcal{D}} \to 0$. Hence a subsequence of $\{u_{\mathcal{D}, \delta t}(\mathbf{x}, 0)\}$, which we still denote by $\{u_{\mathcal{D}, \delta t}(\mathbf{x}, 0)\}$ converges to $u_0(\mathbf{x})$ for a.e. $x \in \Omega$ and

also $\beta(u_{\mathcal{D},\delta t}(\mathbf{x},0)) \to \beta(u_0(\mathbf{x}))$ for a.e. $x \in \Omega$. Since $\beta(u_{\mathcal{D},\delta t}(\mathbf{x},0)) \leqslant \|\beta(u_0(\mathbf{x}))\|_{L^\infty(\Omega)}$ the Lebesgue dominated convergence theorem implies

$$\int_\Omega |\beta(u_0(\mathbf{x})) - \beta(u_{\mathcal{D},\delta t}(\mathbf{x},0))| \, d\mathbf{x} \to 0 \text{ as } h_{\mathcal{D}} \to 0.$$

Thus $T_T^0 \to 0$ as $h_{\mathcal{D}} \to 0$. Next we prove that

$$\sum_{n=1}^{N} \sum_{K \in \mathcal{M}} m(K)(\varphi_K^n - \varphi_K^{n-1})\beta(u_K^n) \to \int_0^T \int_\Omega \beta(u(\mathbf{x},t))\varphi_t(\mathbf{x},t) \, d\mathbf{x}dt \qquad (45)$$

as $h_{\mathcal{D}}$ and $\delta t \to 0$. We define

$$T_T^1 = \sum_{n=1}^{N} \sum_{K \in \mathcal{M}} m(K)(\varphi_K^n - \varphi_K^{n-1})\beta(u_K^n) - \int_0^T \int_\Omega \beta(u(\mathbf{x},t))\varphi_t(\mathbf{x},t) \, d\mathbf{x}dt,$$

and we add and subtract $\int_{t_{n-1}}^{t_n} \int_K \beta(u_K^n)\varphi_t(\mathbf{x},t) \, d\mathbf{x}dt$ in each term to obtain

$$\begin{aligned} T_T^1 &= \sum_{n=1}^{N} \sum_{K \in \mathcal{M}} m(K)\beta(u_K^n) \int_{t_{n-1}}^{t_n} (\varphi_t(\mathbf{x}_K,t) - \varphi_t(\mathbf{x},t)) \, d\mathbf{x}dt \\ &\quad + \int_0^T \int_\Omega (\beta(u_{\mathcal{D},\delta t}(\mathbf{x},t)) - \beta(u(\mathbf{x},t)))\varphi_t(\mathbf{x},t) \, d\mathbf{x}dt. \end{aligned} \qquad (46)$$

We have that for all $x \in K$ and all $K \in \mathcal{M}$ it holds

$$|\varphi_t(\mathbf{x}_K,t) - \varphi_t(\mathbf{x},t)| \leqslant C(h_{\mathcal{D}})$$

where $C(h_{\mathcal{D}}) \to 0$ as $h_{\mathcal{D}} \to 0$. The absolute value of the first term on the right-hand side of (46) is bounded by

$$\begin{aligned} C(h_{\mathcal{D}}) \sum_{n=1}^{N} \sum_{K \in \mathcal{M}} \delta t m(K)|\beta(u_K^n)| &\leqslant C(h_{\mathcal{D}})(Tm(\Omega))^{1/2} (\sum_{n=1}^{N} \sum_{K \in \mathcal{M}} \delta t m(K)(\beta(u_K^n))^2)^{1/2} \\ &\leqslant C(h_{\mathcal{D}})Tm(\Omega)^{1/2}\|\beta(u_{\mathcal{D},\delta t})\|_{L^\infty(0,T;L^2(\Omega))}, \end{aligned}$$

which tends to zero as $h_{\mathcal{D}} \to 0$ in view of the a priori estimate (30). Further, since $|\varphi_t(\mathbf{x},t)| \leqslant C_\varphi$, we can estimate the absolute value of the second term in (46) by

$$\begin{aligned} C_\varphi \int_0^T \int_\Omega |\beta(u_{\mathcal{D},\delta t}(\mathbf{x},t)) - \beta(u(\mathbf{x},t))| \, d\mathbf{x}dt &\leqslant C_\varphi \int_0^T \int_\Omega |\widetilde{\beta}_1(u_{\mathcal{D},\delta t}(\mathbf{x},t)) - \widetilde{\beta}_1(u(\mathbf{x},t))| \, d\mathbf{x}dt \\ &\quad + C_\varphi \int_0^T \int_\Omega |\widetilde{\beta}_2(u_{\mathcal{D},\delta t}(\mathbf{x},t)) - \widetilde{\beta}_2(u(\mathbf{x},t))| \, d\mathbf{x}dt, \end{aligned}$$

where $\widetilde{\beta}_1$ and $\widetilde{\beta}_2$ are given by (32)-(34). Since $u_{\mathcal{D},\delta t} \to u$ strongly in $L^2(Q_T)$, a subsequence of $\{u_{\mathcal{D},\delta t}\}$, which we still denote by $\{u_{\mathcal{D},\delta t}\}$ converges to u a.e. in Ω. The first term on right-hand side of the expression above converges to zero by the Lebesgue dominated convergence theorem. The convergence to zero of the second term can be deduced from the Lipschitz continuity of $\widetilde{\beta}_2$ and the strong convergence of $u_{\mathcal{D},\delta t}$ to u in $L^2(Q_T)$.

6.2 Convection term

Next, we show that

$$E = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} (\varphi_K^{n-1} - \varphi_\sigma^{n-1}) V_{K,\sigma} \overline{u_{K,\sigma}^n} + \int_0^T \int_\Omega u(\mathbf{x},t) \mathbf{V}(\mathbf{x}) \cdot \nabla \varphi(\mathbf{x},t) \, d\mathbf{x} dt \to 0 \quad (47)$$

as $h_\mathcal{D}, \delta t$ tend to zero. As it was done in in the proof of the Lemma 2.4 we write the left-hand side part of (47) as

$$\sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} (\varphi_K^{n-1} - \varphi_\sigma^{n-1}) V_{K,\sigma} \overline{u_{K,\sigma}^n} = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} V_{K,\sigma} (\varphi_K^{n-1} - \varphi_\sigma^{n-1}) u_K^n$$
$$- \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K, V_{K,\sigma} \leqslant 0} V_{K,\sigma} (\varphi_K^{n-1} - \varphi_\sigma^{n-1})(u_K^n - u_\sigma^n), \quad (48)$$

and the following estimate holds

$$| \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K, V_{K,\sigma} \leqslant 0} V_{K,\sigma} (\varphi_K^{n-1} - \varphi_\sigma^{n-1})(u_K^n - u_\sigma^n)| \leqslant h_\mathcal{D} \cdot \|\mathbf{V}\|_{L^\infty(\Omega)} \sum_{n=1}^{N} \delta t |\varphi^{n-1}|_X |u^n|_X$$

$$\leqslant C h_\mathcal{D} \cdot \|\mathbf{V}\|_{L^\infty(\Omega)} \sum_{n=1}^{N} \delta t \|\nabla_\mathcal{D} \varphi^{n-1}\|_{L^2(\Omega)} \|\nabla_D u^n\|_{L^2(\Omega)}$$

$$\leqslant C h_\mathcal{D} \cdot \|\mathbf{V}\|_{L^\infty(\Omega)} \sum_{n=1}^{N} \delta t \|\nabla_\mathcal{D} \varphi^{n-1}\|_{L^2(\Omega)}^2 + C h_\mathcal{D} \cdot \|\mathbf{V}\|_{L^\infty(\Omega)} \|\nabla_{\mathcal{D},\delta t} u_{\mathcal{D},\delta t}\|_{L^2(Q_T)}^2.$$

The second term in the right-hand side of the expression above is bounded be because of the a priori estimate (29) and the first term can be controlled via the consistency of the discrete gradient given by Lemma 2.1 and the regularity of φ; indeed

$$\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \int_\Omega |\nabla_\mathcal{D} \varphi^{n-1}(\mathbf{x})|^2 \, d\mathbf{x} dt \leqslant 3 \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \int_\Omega |\nabla_\mathcal{D} \varphi^{n-1}(\mathbf{x}) - \nabla \varphi(\mathbf{x}, t_{n-1})|^2 \, d\mathbf{x} dt$$

$$+ 3 \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \int_\Omega |\nabla \varphi(\mathbf{x}, t_{n-1}) - \nabla \varphi(\mathbf{x}, t)|^2 \, d\mathbf{x} dt + 3 \int_0^T \int_\Omega |\nabla \varphi(\mathbf{x}, t)|^2 \, d\mathbf{x} dt$$

$$\leqslant (C_\varphi(\delta t) + C h_\mathcal{D}^2) T m(\Omega) + 3 \|\nabla \varphi\|_{L^2(Q_T)}^2 \leqslant C,$$

where $C_\varphi(\delta t)$ tends to zero as $\delta t \to 0$. Thus the second term in the right hand side of (48) tends to zero as $h_\mathcal{D}, \delta t \to 0$. Let us define E_1 and E_2

$$E_1 = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} V_{K,\sigma} (\varphi_K^{n-1} - \varphi_\sigma^{n-1}) u_K^n + \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} u_K^n \int_K \nabla \varphi(\mathbf{x}, t_{n-1}) \cdot \mathbf{V}(\mathbf{x}) \, d\mathbf{x},$$

$$E_2 = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} u_K^n \int_K \nabla \varphi(\mathbf{x}, t_{n-1}) \cdot \mathbf{V}(\mathbf{x}) \, d\mathbf{x} - \int_0^T \int_\Omega u(\mathbf{x}, t) \nabla \varphi(\mathbf{x}, t) \cdot \mathbf{V}(\mathbf{x}) \, d\mathbf{x} dt.$$

so that also in view of (48) $\lim_{h_\mathcal{D}, \delta t \to 0} E = \lim_{h_\mathcal{D}, \delta t \to 0} E_1 - \lim_{h_\mathcal{D}, \delta t \to 0} E_2$. We will successively establish that E_1 and E_2 converges to zero as $h_\mathcal{D}, \delta t \to 0$. To begin with let us remark that integrating by parts in the expression of E_1 yields $E_1 = E_{11} - E_{12}$, where

$$E_{11} = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} V_{K,\sigma} \varphi_K^{n-1} u_K^n - \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} u_K^n \int_K \varphi(\mathbf{x}, t_{n-1}) \nabla \cdot \mathbf{V}(\mathbf{x}) \, d\mathbf{x}$$

and where

$$E_{12} = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} V_{K,\sigma} \varphi_\sigma^{n-1} u_K^n - \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} u_K^n \int_\sigma \varphi(\mathbf{x}, t_{n-1}) \mathbf{V}(\mathbf{x}) \cdot \mathbf{n}_{K,\sigma} d\gamma.$$

We first prove that $\lim_{h_\mathcal{D}, \delta t \to 0} E_{11} = 0$.

$$E_{11} = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} u_K^n \int_K (\varphi_K^{n-1} - \varphi(\mathbf{x}, t_{n-1})) \nabla \cdot \mathbf{V}(\mathbf{x})) \, d\mathbf{x}.$$

in view of regularity of the function φ we obtain

$$|E_{11}| \leqslant C_\varphi h_\mathcal{D} \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} |u_K^n| \int_K |\nabla \cdot \mathbf{V}(\mathbf{x})| \, d\mathbf{x} \leqslant C_\varphi h \int_{Q_T} |u_{\mathcal{D}, \delta t}(\mathbf{x}, t) \nabla \cdot \mathbf{V}(\mathbf{x})| \, d\mathbf{x}$$

Finally applying the Cauchy-Schwarz inequality yields

$$|E_{11}| \leqslant C_\varphi h \|u_{\mathcal{D}, \delta t}\|_{L^2(Q_T)} \|\nabla \cdot \mathbf{V}\|_{L^2(Q_T)}$$

so that $|E_{11}| \to 0$ as $h_\mathcal{D} \to 0$. Next we consider the term E_{12}, which can be written as

$$E_{12} = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} u_K^n \int_\sigma (\varphi_\sigma^{n-1} - \varphi(\mathbf{x}, t_{n-1})) \mathbf{V}(\mathbf{x}) \cdot \mathbf{n}_{K,\sigma} d\gamma.$$

In order to show that $E_{12} \to 0$ as $h_\mathcal{D}, \delta t \to 0$ we first remark that since $\mathbf{n}_{K,\sigma} = -\mathbf{n}_{L,\sigma}$ for any pair of neighbor volumes K, L, and in view of the boundary condition on φ one has that

$$\sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} V_{K,\sigma} \varphi_\sigma^{n-1} u_\sigma^n = 0$$

and also

$$\sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} u_\sigma^n \int_\sigma \varphi(\mathbf{x}, t_{n-1}) \mathbf{V}(\mathbf{x}) \cdot \mathbf{n}_{K,\sigma} d\gamma = 0.$$

Hence, the term E_{12} can be written as

$$E_{12} = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} (u_K^n - u_\sigma^n) \int_\sigma (\varphi_\sigma^{n-1} - \varphi(\mathbf{x}, t_{n-1})) \mathbf{V}(\mathbf{x}) \cdot \mathbf{n}_{K,\sigma} d\gamma.$$

Therefore, in view of the regularity of φ and $\mathbf{V}$ we have that

$$|E_{12}| \leqslant C \max_{\mathbf{x} \in \Omega} \mathbf{V}(\mathbf{x}) \cdot h \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} m(\sigma) |u_K^n - u_\sigma^n|.$$

Applying Cauchy-Schwarz inequality we obtain

$$|E_{12}| \leqslant C d^{\frac{1}{2}} \max_{\mathbf{x} \in \Omega} \mathbf{V}(\mathbf{x}) \cdot h_{\mathcal{D}} (\sum_{n=1}^{N} \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} \delta t m(\sigma) \frac{(u_K^n - u_\sigma^n)^2}{d_{K,\sigma}})^{\frac{1}{2}} \cdot (\sum_{n=1}^{N} \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} \delta t \frac{m(\sigma) d_{K,\sigma}}{d})^{\frac{1}{2}}.$$

In view of Lemma 2.2 we obtain

$$|E_{12}| \leqslant C d \max_{\mathbf{x} \in \Omega} \mathbf{V}(\mathbf{x}) m(\Omega)^{\frac{1}{2}} T^{\frac{1}{2}} \cdot h_{\mathcal{D}} \|\nabla_{\mathcal{D},\delta t} u_{\mathcal{D},\delta t}\|_{L(Q_T)},$$

so that in view of the a priori estimate (29) one has that $|E_{12}| \to 0$ as $h_{\mathcal{D}} \to 0$, so that $E_1 = E_{11} - E_{12} \to 0$ as $h_{\mathcal{D}}, \delta t \to 0$. It remains to prove that E_2 converges to zero.

Adding and subtracting $\int_{t_{n-1}}^{t_n} \int_K u_K^n \nabla \varphi(\mathbf{x}, t) \cdot \mathbf{V}(\mathbf{x}) \, d\mathbf{x} dt$ from each term of E_2, yields

$$\begin{aligned} E_2 \ =& \sum_{n=1}^{N} \sum_{K \in \mathcal{M}} \int_{t_{n-1}}^{t_n} \int_K u_K^n (\nabla \varphi(\mathbf{x}, t_{n-1}) - \nabla \varphi(\mathbf{x}, t)) \cdot \mathbf{V}(\mathbf{x}) \, d\mathbf{x} dt \\ &- \sum_{n=1}^{N} \sum_{K \in \mathcal{M}} \int_{t_{n-1}}^{t_n} \int_K (u(\mathbf{x}, t) - u_{\mathcal{D},\delta t}(\mathbf{x}, t)) \nabla \varphi(\mathbf{x}, t) \cdot \mathbf{V}(\mathbf{x}) \, d\mathbf{x} dt. \end{aligned}$$

Finally, in view of the regularity of φ, the a priori estimate (29), and to the fact that $u_{\mathcal{D},\delta t} \to u$ strongly in $L^2(Q_T)$ we conclude that $|E_2|$ tends to zero as $h_{\mathcal{D}}, \delta t \to 0$.

6.3 Diffusion term

We show below that

$$T_D^1 = \sum_{n=1}^{N} \delta t \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} (\varphi_K^{n-1} - \varphi_\sigma^{n-1}) F_{K,\sigma}(u^n) - \int_0^T \int_\Omega \nabla \varphi(\mathbf{x}, t) \cdot \mathbf{\Lambda}(\mathbf{x}) \nabla u(\mathbf{x}, t) \, d\mathbf{x} dt$$

tends to zero as $h_{\mathcal{D}}, \delta t \to 0$. In view of (23) one has that

$$T_D^1 = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \int_\Omega (\nabla_{\mathcal{D}} \varphi^{n-1} \cdot \mathbf{\Lambda}(\mathbf{x}) \nabla_{\mathcal{D}} u^n - \nabla \varphi(\mathbf{x}, t) \cdot \mathbf{\Lambda}(\mathbf{x}) \nabla u(\mathbf{x}, t)) \, d\mathbf{x} dt.$$

Adding and subtracting the term $\int_{\Omega} \nabla\varphi(\mathbf{x}, t) \cdot \boldsymbol{\Lambda}(\mathbf{x}) \nabla_{\mathcal{D}} u^n \, d\mathbf{x}$ we set T_D^1 in the form $T_D^1 = T_D^2 + T_D^3$ with

$$T_D^2 = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \int_{\Omega} (\nabla_{\mathcal{D}} \varphi^{n-1} - \nabla\varphi(\mathbf{x}, t)) \cdot \boldsymbol{\Lambda}(\mathbf{x}) \nabla_{\mathcal{D}} u^n \, d\mathbf{x} dt$$

and

$$T_D^3 = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \int_{\Omega} \nabla\varphi(\mathbf{x}, t) \cdot \boldsymbol{\Lambda}(\mathbf{x}) (\nabla_{\mathcal{D}} u^n - \nabla u(\mathbf{x}, t)) \, d\mathbf{x} dt.$$

The term T_D^3 tends to zero as $h_{\mathcal{D}}, \delta t \to 0$, since $\nabla_{\mathcal{D}, \delta t} u_{\mathcal{D}, \delta t}$ tends to ∇u weakly in $L^2(Q_T)$. On the other hand the term T_D^2 can be written in the form $T_D^2 = T_D^4 + T_D^5$ with

$$T_D^4 = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \int_{\Omega} (\nabla_{\mathcal{D}} \varphi^{n-1} - \nabla\varphi(\mathbf{x}, t_{n-1})) \cdot \boldsymbol{\Lambda}(\mathbf{x}) \nabla_{\mathcal{D}} u^n \, d\mathbf{x} dt$$

and

$$T_D^5 = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \int_{\Omega} \nabla(\varphi(\mathbf{x}, t_{n-1}) - \varphi(\mathbf{x}, t)) \cdot \boldsymbol{\Lambda}(\mathbf{x}) \nabla_{\mathcal{D}} u^n \, d\mathbf{x} dt.$$

It follows from (29), Lemma 2.1 and the regularity of φ that T_D^4 and T_D^5 tends to zero as $h_{\mathcal{D}}, \delta t \to 0$ and so do T_D^2 and T_D^1.

6.4 Reaction term

Let us show that

$$T_R \to \int_0^T \int_{\Omega} F(u(\mathbf{x}, t)) \varphi(\mathbf{x}, t) \, d\mathbf{x} dt$$

as $h_{\mathcal{D}}$ and k tend to zero. For this purpose, we introduce

$$T_R^1 = \sum_{n=1}^{N} \sum_{K \in \mathcal{M}} \int_{t_{n-1}}^{t_n} \int_K (\varphi_K^{n-1} - \varphi(\mathbf{x}, t)) F(u_K^n) \, d\mathbf{x} dt,$$

$$T_R^2 = \sum_{n=1}^{N} \sum_{K \in \mathcal{M}} \int_{t_{n-1}}^{t_n} \int_K \varphi(\mathbf{x}, t) (F(u_K^n) - F(u(\mathbf{x}, t))) \, d\mathbf{x} dt.$$

We obtain the convergence result similarly as for the time evolution term; more precisely we split the reaction F into a bounded and a Lipschitz continuous parts by setting

$$F_1(s) = \begin{cases} F(s) & 0 \leqslant s \leqslant M \\ 0 & \text{otherwise,} \end{cases} \qquad F_2(s) = \begin{cases} 0 & 0 \leqslant s \leqslant M \\ F(s) & \text{otherwise,} \end{cases}$$

and

$$y(s) = \begin{cases} \dfrac{F(M)}{M}s & 0 \leqslant s \leqslant M \\ 0 & \text{otherwise.} \end{cases}$$

We, then, define $\widetilde{F}_1 = F_1 - y$ and $\widetilde{F}_2 = F_2 + y$ which are both continuous; moreover $|\widetilde{F}_1|$ is bounded by $C_{\widetilde{F}} = \max_{0 \leqslant s \leqslant M} |F(s)| + F(M)$, and $\widetilde{F}_2$ is Lipschitz continuous with Lipschitz constant $L_{\widetilde{F}} = \max(L_F, F(M)/M)$. In view of the regularity of φ one has

$$|\varphi(\mathbf{x}, t) - \varphi_K^{n-1}| \leqslant C(h_{\mathcal{D}} + \delta t) \text{ for all } x \in K, t \in (t_{n-1}, t_n],$$

so that

$$|T_R^1| \leqslant C(h_{\mathcal{D}} + \delta t) \int_0^T \int_\Omega |\widetilde{F}_1(u_{\mathcal{D}, \delta t}(\mathbf{x}, t)) + \widetilde{F}_2(u_{\mathcal{D}, \delta t}(\mathbf{x}, t))| \, d\mathbf{x} dt$$

$$\leqslant C(h_{\mathcal{D}} + \delta t)(C_{\widetilde{F}} T m(\Omega) + L_{\widetilde{F}} T^{\frac{1}{2}} m(\Omega)^{\frac{1}{2}} \|u_{\mathcal{D}, \delta t}\|_{L^2(Q_T)}),$$

which by the a priori estimate (29) implies that $|T_R^1| \to 0$ as $h_{\mathcal{D}}, \delta t \to 0$. Since φ is bounded we can estimate the second term as

$$|T_R^2| \leqslant C \int_0^T \int_\Omega |F(u_{\mathcal{D}, \delta t}(\mathbf{x}, t)) - F(u(\mathbf{x}, t))| \, d\mathbf{x} dt$$

$$\leqslant \int_0^T \int_\Omega C|\widetilde{F}_1(u_{\mathcal{D}, \delta t}(\mathbf{x}, t)) - \widetilde{F}_1(u(\mathbf{x}, t))| \, d\mathbf{x} dt + \int_0^T \int_\Omega C|\widetilde{F}_2(u_{\mathcal{D}, \delta t}(\mathbf{x}, t)) - \widetilde{F}_2(u(\mathbf{x}, t))| \, d\mathbf{x} dt.$$

The convergence is can be proved by applying similar arguments as for the time evolution term.

6.5   Source term

We deduce from the regularity of φ that

$$T_S = \sum_{n=1}^N \sum_{K \in \mathcal{M}} \int_{t_{n-1}}^{t_n} \int_K \varphi(\mathbf{x}_K, t_{n-1}) q(\mathbf{x}, t) \, d\mathbf{x} dt \to \int_0^T \int_\Omega \varphi(\mathbf{x}, t) q(\mathbf{x}, t) \, d\mathbf{x} dt$$

as $h_{\mathcal{D}}, \delta t \to 0$.

6.6   Convergence to a weak solution of Problem $(\mathcal{P})$

In view of Theorem 6.1 $\{u_{\mathcal{D}, \delta t}\}$ strongly converges to u in $L^2(Q_T)$, with $u \in L^2(0, T; H_0^1(\Omega))$, and it follows from (30) that $\beta(u) \in L^\infty(0, T; L^2(\Omega))$. Moreover we deduce from the density of the set Φ in the set $\{\varphi \in L^2(0, T; H_0^1(\Omega)), \varphi_t \in L^\infty(Q_T), \varphi(\cdot, T) = 0\}$ that u is a weak solution of the continuous problem $(\mathcal{P})$ in the sense of Definition 2.1. In the case that F is nondecreasing so that the solution of Problem $(\mathcal{P})$ is unique (cf. Remark 2.1) we conclude that the whole family $\{u_{\mathcal{D}, \delta t}\}$ converges to u.

7 Numerical simulations

In this section we present the results of numerical simulations. The purpose is to test our scheme in the case of problems with a known analytical solution.

7.1 Numerical Test I

We consider the equation

$$\frac{\partial(u + u^{\frac{1}{2}})}{\partial t} - \nabla \cdot (\boldsymbol{\Lambda}(\mathbf{x})\nabla u) + \nabla \cdot (\mathbf{V}(\mathbf{x})u) + \frac{1}{2}u^{\frac{1}{2}} = 0$$

in the 3-dimensional space domain $\Omega = (0,2) \times (0,1) \times (0,1)$. We define the discontinuous $\boldsymbol{\Lambda}$ and $\mathbf{V}$ fields as follows
For all $x_1 \leqslant 1$ we set

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{V} = (4,0,0);$$

for all $x_1 > 1$ we set

$$\boldsymbol{\Lambda} = \begin{pmatrix} 8 & -5 & -2 \\ -5 & 20 & -7 \\ -2 & -7 & 19 \end{pmatrix} \quad \text{and} \quad \mathbf{V} = (4,7,7).$$

The initial and the Dirichlet boundary conditions are given by the exact solution

$$u(\mathbf{x}, t) = e^{x_1 + x_2 + x_3 - t - 3}.$$

We remark that the velocity field $\mathbf{V}$ and the total flux $\boldsymbol{\Lambda}(\mathbf{x})\nabla u + \mathbf{V}(\mathbf{x})u$ have a continuous normal trace across the discontinuity $x = 1$. We perform the simulations on 3-dimensional hexahedral meshes with random refinement (see Figure 1), so that the mesh is nonmatching. In Table 1 we present simulation results with various mesh sizes $h_{\mathcal{D}}$ and time steps k; we denote by Err the maximum relative error in L^2-norm, namely

$$Err = \max_{n \in \{1, \dots, N\}} \frac{\|u_{h,t}(\cdot, t_n) - u(\cdot, t_n)\|_{L^2(\Omega)}}{\|u(\cdot, t_n)\|_{L^2(\Omega)}}.$$

7.2 Numerical test II

We consider a degenerate parabolic equation which possesses a traveling wave solution, namely

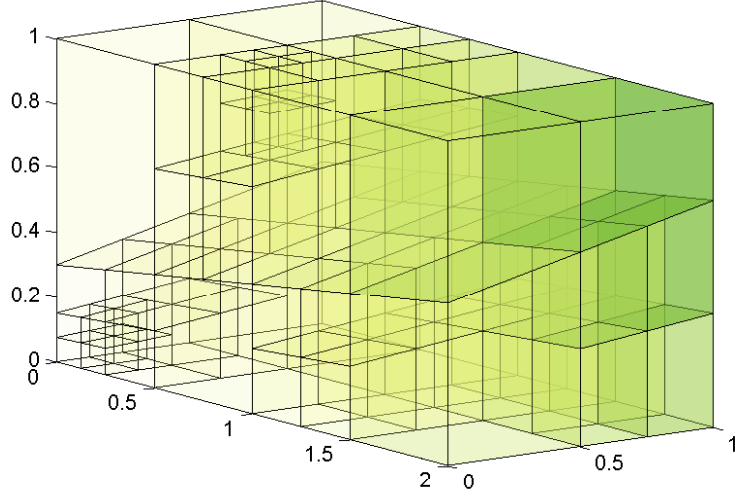$$\frac{\partial(u^{\frac{1}{2}})}{\partial t} - \nabla \cdot (\delta \nabla u) + \nabla \cdot ((v,0,0)u) = 0$$

Figure 1: Approximate solution on the nonmatching hexahedral mesh at $t = 1$

N	h	# of elements	# of faces	Err
50	0.75	165	672	0.03575
100	0.375	837	3324	0.01432
200	0.1875	3203	11550	0.00648
400	0.0938	18533	60633	0.00305

Table 1: Number of time steps N, mesh diameter $h_{\mathcal{D}}$, number of elements, number of faces and the relative error for nonmatching hexahedral meshes

in the domain

$$\Omega = (0,1)^3 \text{ and } T = 1.$$

This equation admits the following 1-dimensional exact solution

$$u(x,y,t) = (1 - e^{\frac{v}{2\delta}(x-vt-p)})^2 \text{ for } x \leqslant vt + p,$$

$$u(x,y,t) = 0 \text{ for } x > vt + p$$

where p, v, δ are parameters still to be defined. We set $p = 0.2$, $v = 0.8$, and consider two values of δ, namely 0.01, 0.0001. The initial state is given by the exact solution at the time $t = 0$ and we prescribe corresponding Dirichlet boundary conditions on the sides $x = 0$ and $x = 1$. The null flux boundary condition is imposed on the remaining part of the boundary.

Since the scheme does not preserve the maximum principle, it is necessary to define the function $\beta(u) = u^{\frac{1}{2}}$ for negative values as well, which leads us to set $\beta(-u) = -\beta(u)$. Further one

has to solve the system of nonlinear equations

$$
\begin{cases}
m(K)(\beta(u_K^n) - \beta(u_K^{n-1})) + k \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}(u^n) \\
\qquad + k \sum_{\sigma \in \mathcal{E}_K} V_{K,\sigma} \overline{u_{K,\sigma}^n} = \delta t \, m(K) q_K^n, \quad \text{for all } K \in \mathcal{M}, \\
(F_{K,\sigma}(u^n) + V_{K,\sigma} \overline{u_{K,\sigma}^n}) + (F_{L,\sigma}(u^n) + V_{L,\sigma} \overline{u_{L,\sigma}^n}) = 0, \quad \text{for all } \sigma \in \mathcal{E}_{int} \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad u_\sigma^n = 0, \quad \text{for all } \sigma \in \mathcal{E}_{ext}.
\end{cases}
\tag{49}
$$

Since $\beta'(0) = +\infty$ the Newton method can not be directly applied. In order to overcome this difficulty we introduce new discrete unknowns

$$
w^n = \beta(u^n), \text{ and thus } u^n = \varphi(w^n), \text{ where } \varphi = \beta^{-1}.
$$

In view of (10) and (25) the nonlinear system becomes

$$
\begin{cases}
m(K)(w_K^n - w_K^{n-1}) + \delta t \sum_{\sigma, \sigma' \in \mathcal{E}_K} A^{\sigma\sigma'} (\varphi(w_K^n) - \varphi(w_{\sigma'}^n)) \\
\qquad + \delta t \sum_{\sigma \in \mathcal{E}_K} (V_{K,\sigma}^+ \varphi(w_K^n) + V_{K,\sigma}^- \varphi(w_\sigma^n)) = \delta t \, m(K) q_K^n, \quad \text{for all } K \in \mathcal{M}, \\
\sum_{\sigma' \in \mathcal{E}_K} A^{\sigma\sigma'} (\varphi(w_K^n) - \varphi(w_{\sigma'}^n)) + (V_{K,\sigma}^+ \varphi(w_K^n) + V_{K,\sigma}^- \varphi(w_\sigma^n)) \\
+ \sum_{\sigma' \in \mathcal{E}_L} A^{\sigma\sigma'} (\varphi(w_L^n) - \varphi(w_{\sigma'}^n)) + (V_{L,\sigma}^+ \varphi(w_L^n) + V_{L,\sigma}^- \varphi(w_\sigma^n)) = 0, \quad \text{for all } \sigma \in \mathcal{E}_{int}, \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \varphi(w_\sigma^n) = 0, \quad \text{for all } \sigma \in \mathcal{E}_{ext}.
\end{cases}
\tag{50}
$$

The system (50) depends on $(w_\sigma^n)_{\sigma \in \mathcal{E}}$ only through the terms $(\varphi(w_\sigma^n))_{\sigma \in \mathcal{E}}$. This lead us to choose the discrete unknowns

$$
w_K^n = \beta(u_K^n) \text{ for all } K \in \mathcal{M} \text{ and } u_\sigma^n \text{ for all } \sigma \in \mathcal{E},
$$

so that the system (50) takes the form

$$
\begin{cases}
m(K)(w_K^n - w_K^{n-1}) + \delta t \sum_{\sigma, \sigma' \in \mathcal{E}_K} A^{\sigma\sigma'} (\varphi(w_K^n) - u_{\sigma'}^n) \\
\qquad + \delta t \sum_{\sigma \in \mathcal{E}_K} (V_{K,\sigma}^+ \varphi(w_K^n) + V_{K,\sigma}^- u_\sigma^n) = \delta t \, m(K) q_K^n, \quad \text{for all } K \in \mathcal{M}, \\
\sum_{\sigma' \in \mathcal{E}_K} A^{\sigma\sigma'} (\varphi(w_K^n) - u_{\sigma'}^n) + (V_{K,\sigma}^+ u_K^n + V_{K,\sigma}^- u_\sigma^n) \\
+ \sum_{\sigma' \in \mathcal{E}_L} A^{\sigma\sigma'} (\varphi(w_L^n) - u_{\sigma'}^n) + (V_{L,\sigma}^+ \varphi(w_L^n) + V_{L,\sigma}^- u_\sigma^n) = 0, \quad \text{for all } \sigma \in \mathcal{E}_{int}, \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad u_\sigma^n = 0, \quad \text{for all } \sigma \in \mathcal{E}_{ext}.
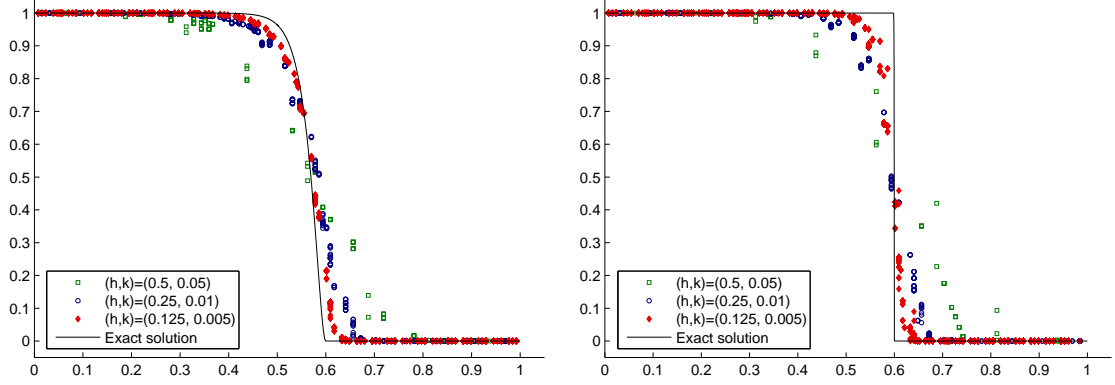\end{cases}
\tag{51}
$$

34

Figure 2: The approximate solution profiles at the time $t = 0.5$ for $\delta = 0.01$ and $\delta = 0.0001$

Remark 7.1. *We remark that the nonlinear system (49) (or a linear one arising during the Newton's procedure) has a special structure; more specifically, for each $K \in \{1, \ldots, card(\mathcal{M})\}$ the K-th equation does not contain any unknown different from u_K and $(u_\sigma)_{\sigma \in \mathcal{E}_K}$ (here we denote by K both the control volume and the index of the unknown u_K); therefore one can algebraically eliminate u_K, so that the number of equations to solve becomes $card(\mathcal{E})$.*

Since we do not impose many constraints on the mesh (in particular it can be nonconforming), it is not difficult to perform a local grid refinement. Finally note that there is a possibility to reduce the number of unknowns by using a method introduced in [14]; one can eliminate the interior interface unknowns $(u_\sigma)_{\sigma \in \mathcal{E}_{int}}$ by expressing them as a consistent barycentric combinations of the values u_K.

References

[1] R. A. Adams, J.F. Fournier, Sobolev Spaces, *Pure and Applied Mathematics*, vol. 140, Academic Press, New York-London, 2003.

[2] M. Afif, B. Amaziane, Convergence of finite volume schemes for a degenerate convection-diffusion equation arising in flow in porous media, *Comput. Methods Appl. Mech. Engrg*, 191, 2002, 5265–5286.

[3] O. Angelini, C. Chavant, E. Chenier, R. Eymard, A finite volume scheme for diffusion problems on general meshes applying monotony constraints, *SIAM J. Numer. Anal*, 47, 2010, 4193–4213.

[4] P. Angot, V. Dolejsi, M. Feistauer, J. Felcman, Analysis of a Combined Barycentric Finite Volume - Nonconforming Finite Element Method for Nonlinear Convection - Diffusion Problem, *Applications of Mathematics*, 43, 1998, 263–310.

[5] T. Arbogast, M. F. Wheeler, N. Zhang, A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media, *SIAM J. Numer. Anal.*, 33, 1996, 1669-1687.

[6] L.A. Baughman, N.J. Walkington, Co-volume methods for degenerate parabolic problems, *Numer. Math*, 64, 1993, 45–67.

[7] Z. Chen, R.E. Ewing, E.Q. Jiang, A.M. Spagnuolo, Error analysis for characteristics-based methods for degenerate parabolic problems, *SIAM J. Numer. Anal.* 40, 2002, 1491–1515.

[8] Y. Coudire, J.-P. Vila, Ph. Villedieu, Convergence rate of a finite volume scheme for a two dimensional diffusion convection problem, *M2AN Math. Model. Numer. Anal.*, 33, 1999, 493–516.

[9] C.N. Dawson, Analysis of an upwind-mixed finite element method for nonlinear contaminant trans- port equations, *SIAM J. Numer. Anal.*, 35, 1998, 1709–1724.

[10] C. Dawson, V. Aizinger, Upwind mixed methods for transport equations, *Comput. Geosci*, 3, 1999, 93–110.

[11] K. Deimling, Nonlinear Functional Analysis, *Springer-Verlag*, Berlin-Heidelberg 1985.

[12] J. Droniou, R. Eymard, T. Gallout, and R. Herbin, A unified approach to Mimetic Finite Difference, Hybrid Finite Volume and Mixed Finite Volume methods, *Math. Models Methods Appl. Sci*, 20, 2010, 265–295.

[13] R. Eymard, T. Gallouët, R. Herbin, Finite Volume Methods, *Handbook of Numerical Analysis*, vol. 7, P.G. Ciarlet and J.L. Lions eds Elsevier Science B.V., Amsterdam, 2000.

[14] R. Eymard, T. Gallouët, R. Herbin, Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI: a scheme using stabilization and hybrid interfaces, *to appear in IMA J. of Num. Anal.*.

[15] R. Eymard, T. Gallouët, R. Herbin, and A. Michel, Convergence of a finite volume scheme for nonlinear degenerate parabolic equations, *Numer. Math.*, 92, 2002, 41–82.

[16] R. Eymard, M. Gutnic, D. Hilhorst, The finite volume method for the Richards equation, *Comput. Geosci.*, 3, 2000, 259–294.

[17] R. Eymard, D. Hilhorst, M. Vohralík, A combined finite volume–nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems, *Numer. Math.*, 105, 2006, 73–131.

[18] R. Eymard, D. Hilhorst, M. Vohralík, A combined finite volume-finite element scheme for the discretization of strongly nonlinear convection-diffusion-reaction problems on non-matching grids, *Numer. Methods for Partial Differ. Equations*, 26, 2009, 612–646.

[19] M. Feistauer, J. Felcman, M. Lukacova-Medvidova, On the Convergence of a Combined Finite Volume-Finite Element Method for Nonlinear Convection-Diffusion Problems, *Numer. Methods for Partial Differ. Equations*, 13, 1997, 163–190.

[20] R. Herbin and F. Hubert, Benchmark on discretization schemes for anisotropic diffusion problems on general grids for anisotropic heterogeneous diffusion problems, *Finite Volumes for Complex App. V*, 2008, 659–692.

[21] J. Kačur, Solution of Degenerate Convection-Diffusion Problems by the Method of Characteristics, *SIAM J. Numer. Anal.*, 39, 2001, 858–879.

[22] J. Kačur, R. van Keer, Solution of contaminant transport with adsorption in porous media by the method of characteristics, *ESAIM: M2AN Math. Model. Numer. Anal.*, 35, 2001, 981–1006.

[23] K. H. Karlsen, N. H. Risebro, J. D. Towers, Upwind difference approximations for degenerate parabolic convection-diffusion equations with a discontinuous coefficient, *IMA J. Numer. Anal.*, 22, 2002, 623–664.

[24] P. Knabner, F. Otto, Solute transport in porous media with equilibrium and nonequilibrium multiple-site adsorption: uniqueness of weak solutions, *Nonlinear Anal.*, 42, 2000, 381–403.

[25] R. H. Nochetto, A. Schmidt, C. Verdi, A posteriori error estimation and adaptivity for degenerate parabolic problems, *Math. Comput.*, 69, 2000, 1–24.

[26] P.H. Nochetto, C. Verdi, Approximation of degenerate parabolic problems using numerical integration, *SIAM J. Numer. Anal.*, 25, 1988, 784–814.

[27] J. Rulla, N. J. Walkington, Optimal rates of convergence for degenerate parabolic problems in two dimensions, *SIAM J. Numer. Anal.*, 33, 1996, 56–67.