Convergence of nonlinear numerical approximations for an elliptic linear problem with irregular data

Robert Eymard, David Maltese

ESAIM: M2AN 55 (2021) 3043–3089
https://doi.org/10.1051/m2an/2021079

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

CONVERGENCE OF NONLINEAR NUMERICAL APPROXIMATIONS FOR AN ELLIPTIC LINEAR PROBLEM WITH IRREGULAR DATA

Robert Eymard* and David Maltese

Abstract. This work is devoted to the study of the approximation, using two nonlinear numerical methods, of a linear elliptic problem with measure data and heterogeneous anisotropic diffusion matrix. Both methods show convergence properties to a continuous solution of the problem in a weak sense, through the change of variable $u = \psi(v)$, where ψ is a well chosen diffeomorphism between $(-1, 1)$ and $\mathbb{R}$, and v is valued in $(-1, 1)$. We first study a nonlinear finite element approximation on any simplicial grid. We prove the existence of a discrete solution, and, under standard regularity conditions, we prove its convergence to a weak solution of the problem by applying Hölder and Sobolev inequalities. Some numerical results, in 2D and 3D cases where the solution does not belong to $H^1(\Omega)$, show that this method can provide accurate results. We then construct a numerical scheme which presents a convergence property to the entropy weak solution of the problem in the case where the right-hand side belongs to L^1. This is achieved owing to a nonlinear control volume finite element (CVFE) method, keeping the same nonlinear reformulation, and adding an upstream weighting evaluation and a nonlinear p-Laplace vanishing stabilisation term.

Mathematics Subject Classification. 65N12, 65N30.

Received January 10, 2021. Accepted November 25, 2021.

1. Introduction

This paper is devoted to the numerical approximation of a second order linear elliptic equation in divergence form with coefficients in $L^\infty(\Omega)$ and measure data. More precisely we consider the following problem: find a function $\overline{u}$ defined on Ω such that

$$-\mathrm{div}(\Lambda \nabla \overline{u}) = f \text{ in } \Omega, \tag{1.1}$$

supplemented with the boundary condition

$$\overline{u} = 0 \text{ on } \partial\Omega, \tag{1.2}$$

under the following assumptions:

- $\Omega \subset \mathbb{R}^d$ $(d \geq 2)$, is a polytopal bounded open set (polygonal if $d = 2$, polyhedral if $d \geq 3$), (1.3a)
- $\Lambda \in L^\infty(\Omega)^{d\times d}$ and there exists $\underline{\lambda}, \overline{\lambda} > 0$ such that, for *a.e.* $x \in \Omega$,
  $\Lambda(x)$ is symmetric and, for all $\xi \in \mathbb{R}^d$, $\underline{\lambda}|\xi|^2 \leq \Lambda(x)\xi \cdot \xi \leq \overline{\lambda}|\xi|^2$, (1.3b)

Keywords and phrases. Elliptic equations with irregular data, finite elements, control-volume finite elements, entropy solution.

LAMA, Univ. Gustave Eiffel, Univ. Paris Est Créteil, CNRS, F-77454 Marne-la-Vallée, France.
*Corresponding author: robert.eymard@univ-eiffel.fr

– $f \in M(\Omega)$, where we denote by $M(\Omega)$ the space of Radon measures, defined as the topological dual space of $C(\overline{\Omega})$, the space of continuous functions on $\overline{\Omega}$ with its usual norm. (1.3c)

This type of problem may arise for example in some models of underground oil or water resources management. Then f is the source term of the quantity diffused by the system. In the case where it is possible to consider $f \in L^2(\Omega)$, the natural framework of the problem consists in considering the standard weak formulation of (1.1) in $H_0^1(\Omega)$. But in the case where f is modeling a singular source term at the scale of the domain, the right-hand side can be a measure instead of a function. For example, in underground fluid management [19], the wells, which are used for injection or production purposes, are cylindrical holes with very small diameters compared to the size of the domain. Therefore, the source terms are accurately modeled by measures supported by lines.

A proper mathematical sense to a solution of problems (1.1) and (1.2) under Assumptions 1.3 can be given as follows.

Definition 1.1. We define the space $\mathcal{S}_d(\Omega)$ containing any solution and the space $\mathcal{T}_d(\Omega)$ containing the test functions by

$$\mathcal{S}_d(\Omega) = \bigcap_{r\in\left(1,\frac{d}{d-1}\right)} W_0^{1,r}(\Omega) \text{ and } \mathcal{T}_d(\Omega) = \bigcup_{r\in(d,+\infty)} W_0^{1,r}(\Omega) \subset C(\overline{\Omega}). \tag{1.4}$$

We say that a function $\overline{u}$ defined over Ω is a weak solution of problems (1.1) and (1.2) if

$$\overline{u} \in \mathcal{S}_d(\Omega) \text{ and } \int_\Omega \Lambda\nabla\overline{u}\cdot\nabla w\,\mathrm{d}x = \int_\Omega w\,\mathrm{d}f, \text{ for any } w \in \mathcal{T}_d(\Omega), \tag{1.5}$$

where we denote by $\int_\Omega w\,\mathrm{d}f$ the quantity $\langle f, w\rangle_{M(\Omega),C(\overline{\Omega})}$.

The existence of a weak solution for any $d \geq 2$ is given in [4]. Its uniqueness is proved for $d = 2$ in [20] for general diffusion fields: the proof relies on a regularity result [22] which holds on domains Ω with C^2 boundary, extended in [21] to all domains with Lipschitz boundaries. In the case $d \geq 3$, this uniqueness result remains true if Λ is sufficiently regular (see [15] for such a proof when $\Lambda = \mathrm{Id}$), but it is no longer true for general diffusion fields (as the ones introduced in the numerical examples in Sect. 4, inspired by the counter-example introduced by Serrin [24] and detailed by Prignet [23]).

This paper is focused on the approximation of problem (1.5). Consistently with the space $\mathcal{S}_d(\Omega)$ introduced in Definition 1.1, we investigate weak/strong convergence of approximate solutions in $W^{1,r}(\Omega)$ for $r \in \left(1, \frac{d}{d-1}\right)$. Let us cite a few different works providing such convergence results.

In [15], a finite volume method is used. It relies on the Two-Point Flux Approximation in grids satisfying an orthogonality condition, which is restricting the kind of meshes which can be used and the kind of anisotropy that can be considered for Λ. The convergence of the approximate solution to the unique weak solution of the problem is proved in the case $\Lambda = \mathrm{Id}$. The method of proof implies the use of nonlinear functions of the discrete unknown as test function, which is easily managed by the use of Two-Point Flux Approximation (see [18] for a discussion about this problem). Since our goal is to consider general diffusion fields Λ, this scheme no longer applies.

The standard finite element framework is considered in [8]: it consists in computing the solution $u_\mathcal{T}$ to the following linear discrete problem

$$u_\mathcal{T} \in \mathcal{V}_\mathcal{T}(\Omega) \text{ and } \int_\Omega \Lambda_\mathcal{T}\nabla u_\mathcal{T}\cdot\nabla w\,\mathrm{d}x = \int_\Omega w\,\mathrm{d}f, \text{ for any } w \in \mathcal{V}_\mathcal{T}(\Omega), \tag{1.6}$$

where $\mathcal{V}_\mathcal{T}(\Omega)$ is the finite dimensional space resulting from the use of the P^1 finite elements on triangles or tetrahedra, and $\Lambda_\mathcal{T}$ is a piecewise constant approximation of Λ. This solution can always be computed, since $\mathcal{V}_\mathcal{T}(\Omega) \subset C(\overline{\Omega})$. But the study of its convergence cannot be done in the classical way, which consists in taking $w = u_\mathcal{T}$ for getting an estimate. Indeed, for $d > 1$, this method does not yield an estimate of $\|u_\mathcal{T}\|_{C(\overline{\Omega})}$, which

would be necessary for passing to the limit. Nevertheless, the convergence of the approximate solution to a weak solution (which is moreover the renormalised solution or equivalently the entropy solution in the sense of Def. 3.1) can be proved in the case where the finite element scheme is similar to that used in the finite volume framework, that is when it relies on a Two-Point Flux interpretation of the finite element scheme. This requires the use of P^1 finite elements on triangles or tetrahedra which satisfy strong geometrical constraints (in 2D, for two triangles sharing the same edge, the sum of the two opposite angles must be lower than π in the case $\Lambda = \mathrm{Id}$). The assumption of general diffusion fields Λ only satisfying Assumption (1.3b) is incompatible with this method of proof.

We therefore propose in this paper two nonlinear numerical methods for the approximation of this problem, with different goals for the two methods. The first one concerns an accurate nonlinear finite element scheme, for which a proof of convergence to a weak solution remains available on general simplicial meshes and fields Λ. The second one concerns a nonlinear Control-Volume Finite-Element scheme (CVFE for short) for which, in addition, a proof of convergence to the entropy weak solution holds in the case where $f \in L^1(\Omega)$.

Notes applying to the whole paper

- We denote by $\|\cdot\|_r$ the norm of $L^r(\Omega)$ for any $r \geq 1$.
- The Lebesgue measure of any subset Q of $\mathbb{R}^d$ is denoted by $|Q|$.
- We use several times the Sobolev inequalities provided by Lemma A.1, involving the coefficients $C_{\mathrm{sob}}^{(r,q)}$.

2. Study of the nonlinear finite element approximation

2.1. Motivation and organisation

Section 2 of this paper explores a non-linear finite element formulation of the problem such that, when taking the solution of the scheme as test function, we get an estimate leading to a convergence property to some function $\overline{u} \in \mathcal{S}_d(\Omega)$. Hence we introduce a strictly increasing diffeomorphism $\psi : (-1,1) \to \mathbb{R}$, and we replace the function $\overline{u}$ by the function $\psi(v)$ where $v(\Omega) \subset (-1,1)$. A similar idea, used in [5] for getting such an estimate, relies on the functions ψ under the form $v \mapsto \left(\frac{1}{(1-|v|)^{1/m}} - 1\right)\mathrm{sign}(v)$ for any $m > 0$, where $\mathrm{sign}(s) = 1$ if $s \geq 0$ and -1 if $s < 0$. Unfortunately, there is no fixed value for m providing the estimate $\overline{u} \in W_0^{1,r}(\Omega)$ for all $r \in (1, d/(d-1))$: one has to let $m \to 0$ for covering the whole desired range of values for r. So this choice cannot be kept in our framework.

Let us sketch the computations which lead to the expression of ψ chosen in this paper. We first transform the problem (1.5) into the problem, formally given as follows: find a function $v : \Omega \to (-1,1)$, such that, for any regular function w vanishing at the boundary,

$$\int_\Omega \psi'(v)\Lambda\nabla v \cdot \nabla w \,\mathrm{d}x = \int_\Omega w \,\mathrm{d}f. \tag{2.1}$$

Assuming that v can be taken as a test function, and denoting for all $s \in (-1,1)$ by $\beta(s) = \int_0^s \sqrt{\psi'(t)}\,\mathrm{d}t$, the following relation holds

$$\underline{\lambda}\|\nabla\beta(v)\|_2^2 = \underline{\lambda}\int_\Omega \psi'(v)|\nabla v|^2 \,\mathrm{d}x \leq \int_\Omega \psi'(v)\Lambda\nabla v \cdot \nabla v \,\mathrm{d}x = \int_\Omega v \,\mathrm{d}f \leq \|f\|_{M(\Omega)}.$$

From this relation, using Hölder and Sobolev inequalities, an estimate on $\|\nabla\psi(v)\|_r$, for all $r \in (1, d/(d-1))$, is derived under the condition that, up to a quantity that can be controlled, β' is bounded by β.

This leads to the following definition for ψ (see Fig. 1 for a graphical representation of ψ, ψ' and β):

$$\psi : \left| \begin{array}{l} (-1,1) \longrightarrow \mathbb{R} \\ s \longmapsto \left(\exp\dfrac{|s|}{1-|s|} - 1\right)\mathrm{sign}(s), \text{ where } \mathrm{sign}(s) = 1 \text{ if } s \geq 0 \text{ and } -1 \text{ if } s < 0. \end{array} \right. \tag{2.2}$$

Remark 2.1. In fact, the more general definition

$$\psi_\alpha : \left| \begin{array}{l} (-1,1) \longrightarrow \mathbb{R} \\ s \longmapsto \left(\exp \dfrac{\alpha|s|}{1-|s|} - 1\right)\mathrm{sign}(s), \end{array} \right. \tag{2.3}$$

for a given $\alpha > 0$, is considered in the numerical section. Since choosing values $\alpha \neq 1$ does not change the methods of the mathematical analysis, we mainly let $\alpha = 1$ in this paper in order to avoid additional notations.

Note that for any $s \in (-1,1)$, we have

$$\psi'(s) = \frac{1}{(1-|s|)^2} \exp \frac{|s|}{1-|s|},$$

and the reciprocal function to ψ is the function $\psi^{-1} : \mathbb{R} \to (-1,1)$ defined for any $t \in \mathbb{R}$ by

$$\psi^{-1}(t) = \frac{\log(1+|t|)}{1+\log(1+|t|)}\mathrm{sign}(t).$$

Note also that this function satisfies $|(\psi^{-1})'(t)| \leq 1$ for any $t \in \mathbb{R}$. It is noticeable that the function $t \mapsto \log(1+|t|)$ also plays an important role in the analysis done in [14] or [15].

It is then possible to define a finite element formulation of the nonlinear problem (2.1), and to prove its convergence to a weak solution of the problem in the sense of Definition 1.1, hence providing a first proof for the convergence of a numerical scheme, using a general simplicial mesh, for a general diffusion field and an irregular right-hand side. Note that we are not able to prove, in the case where $f \in L^1(\Omega)$, the convergence of this scheme to the entropy weak solution in the sense of Definition 3.1 given below, contrary to the modified scheme studied in Section 3. But, as shown in Section 4, this modification leads to a severe loss of numerical convergence order.

Section 2 is organized as follows. In Section 2.2, we present the numerical scheme and Section 2.3 provides the main results that we are able to prove on this scheme. In Section 2.4, estimates of the discrete solution are established, following the ideas sketched above. These estimates allow, in Section 2.5, to prove the existence of at least one solution to the scheme, using the now classical topological degree arguments. Section 2.6 is devoted to the convergence proof of a subsequence of discrete solutions to a weak solution of the problem (recall that the uniqueness of this solution holds only if $d = 2$). The proof of the weak convergence of a sequence of approximate solutions is based first on the compactness of the sequence of approximate solutions and then on the identification of the limit.

2.2. The numerical scheme

Elements and nodes

We define a simplex in dimension $d \geq 1$ as the interior of the convex hull of a given set of $d+1$ points (called its vertices) which are not all contained in the same hyperplane (a simplex is a triangle if $d = 2$ and a tetrahedron if $d = 3$).

We consider a finite family of simplices $\mathcal{T}$ (called the mesh of the domain), which satisfies the following properties.

(1) The family containing all the vertices of the elements of the mesh, called the family of the nodes of the mesh, is denoted by $(z_i)_{i\in\mathcal{N}}$, and, for an element $K \in \mathcal{T}$ the family of the $d+1$ vertices of K is denoted by $(z_i)_{i\in\mathcal{N}_K}$, with $\mathcal{N}_K \subset \mathcal{N}$. The set $\mathcal{N}$ is partitioned into $\mathcal{N} = \mathcal{N}_{\rm int} \cup \mathcal{N}_{\rm ext}$, where for all $i \in \mathcal{N}_{\rm ext}$, $z_i \in \partial\Omega$ (the exterior nodes) and for all $i \in \mathcal{N}_{\rm int}$, $z_i \in \Omega$ (the interior nodes).
(2) The union of the closure of all the elements of $\mathcal{T}$ is equal to $\overline{\Omega}$.
(3) The intersection of two different elements of $\mathcal{T}$ is empty.

(4) For any $K \in \mathcal{T}$, and for any subset $\{i_1, \dots, i_d\}$ of distinct elements of $\mathcal{N}_K$, then, either $\{z_{i_1}, \dots, z_{i_d}\} \subset \partial\Omega$, or there exists a unique $L \in \mathcal{T}$ different from K such that $\{i_1, \dots, i_d\} \subset \mathcal{N}_L$ (which means that neighboring elements share a complete common face).

Such a family $\mathcal{T}$ is then a regular simplicial mesh of Ω in the usual sense of the finite element literature [9].

For $i \in \mathcal{N}$, one denotes by $\mathcal{T}_i$ the set of all $K \in \mathcal{T}$ such that $i \in \mathcal{N}_K$.

For any $K \in \mathcal{T}$, we denote by $|K|$ the measure in $\mathbb{R}^d$ of K, by h_K the diameter of K and by ρ_K the diameter of the largest ball included in K. Then, we define the mesh size $h_{\mathcal{T}}$ and the mesh regularity $\theta_{\mathcal{T}}$ by

$$h_{\mathcal{T}} = \max_{K \in \mathcal{T}} h_K, \qquad \theta_{\mathcal{T}} = \max_{K \in \mathcal{T}} \frac{h_K}{\rho_K}.$$

P^1 basis functions and barycentric coordinates

We denote, for any $i \in \mathcal{N}$, by φ_i the continuous function defined on $\overline{\Omega}$ which is piecewise affine in each $K \in \mathcal{T}$, continuous, and such that $\varphi_i(z_i) = 1$ and $\varphi_i(z_j) = 0$ for all $j \in \mathcal{N} \setminus \{i\}$. Recall that, for any $K \in \mathcal{T}$ and $x \in K$, $(\varphi_i(x))_{i \in \mathcal{N}_K}$ is the family of the barycentric coordinates of point x with respect to the vertices $(z_i)_{i \in \mathcal{N}_K}$ of K.

Approximation space and scheme

We denote by $\mathcal{V}_{\mathcal{T}}(\Omega)$ the conforming finite element space defined by

$$\mathcal{V}_{\mathcal{T}}(\Omega) = \operatorname{span}(\varphi_i)_{i \in \mathcal{N}_{\mathrm{int}}}.$$

We approximate Λ by some piecewise constant function $\Lambda_{\mathcal{T}} : \Omega \to \mathbb{R}^{d \times d}$ such that

$$\Lambda_{\mathcal{T}}(x) = \sum_{K \in \mathcal{T}} \Lambda_K 1_K(x), \tag{2.4}$$

where for any $K \in \mathcal{T}$, $\Lambda_K \in \mathbb{R}^{d \times d}$ is assumed to be symmetric and to satisfy

$$\underline{\lambda}|\xi|^2 \le \Lambda_K \xi \cdot \xi \le \overline{\lambda}|\xi|^2 \text{ for any } \xi \in \mathbb{R}^d. \tag{2.5}$$

The value Λ_K can be chosen as the mean value of $\Lambda_{\mathcal{T}}$ on K, but it can also be chosen as the value of Λ at any point of K if Λ is sufficiently regular (which is done in the numerical examples of Sect. 4).

We then consider the following approximation of problems (1.1) and (1.2).

Definition 2.2. We say that $v_{\mathcal{T}}$ is a solution of the numerical scheme if $v_{\mathcal{T}} \in \mathcal{V}_{\mathcal{T}}(\Omega)$, $\max_{x \in \overline{\Omega}} |v_{\mathcal{T}}(x)| < 1$ and

$$\int_\Omega \psi'(v_{\mathcal{T}}) \Lambda_{\mathcal{T}} \nabla v_{\mathcal{T}} \cdot \nabla w \, \mathrm{d}x = \int_\Omega w \, \mathrm{d}f, \text{ for any } w \in \mathcal{V}_{\mathcal{T}}(\Omega). \tag{2.6}$$

We remark that, since $\Lambda_{\mathcal{T}}$, $\nabla v_{\mathcal{T}}$ and ∇w are constant in each element of $\mathcal{T}$, this numerical scheme leads to the computation, for each $K \in \mathcal{T}$, of the quantity $\int_K \psi'(v_{\mathcal{T}}) \, \mathrm{d}x$. We provide in Section A.1 the mathematical computation of this quantity.

2.3. Main results of Section 2

The first main result of this section is the existence of a solution to the nonlinear scheme.

Theorem 2.3. *There exists (at least) one solution $v_{\mathcal{T}} \in \mathcal{V}_{\mathcal{T}}(\Omega)$ to Scheme (2.6) (in the sense of Def. 2.2).*

Once we have a discrete solution $v_{\mathcal{T}} \in \mathcal{V}_{\mathcal{T}}(\Omega)$ at hand for all meshes, then we can study the convergence of the scheme when the discretisation parameter $h_{\mathcal{T}}$ tends to zero. More precisely, consider a sequence $\left(\mathcal{T}^{(m)}\right)_{m \ge 1}$ of meshes of Ω in the sense specified in Section 2.2 such that

$$h_{\mathcal{T}^{(m)}} \underset{m \to \infty}{\longrightarrow} 0, \tag{2.7}$$

and such that the sequence of regularity is bounded: there exists $\theta^\star$ such that

$$\theta_{\mathcal{T}^{(m)}} \leq \theta^\star, \text{ for any } m \geq 1. \tag{2.8}$$

Let $(\Lambda_{\mathcal{T}^{(m)}})_{m\geq 1}$ be such that (2.4) and (2.5) hold and such that

$$\Lambda_{\mathcal{T}^{(m)}} \underset{m\to\infty}{\longrightarrow} \Lambda \text{ in } L^r(\Omega)^{d\times d} \text{ for any } r \in [1,+\infty). \tag{2.9}$$

Note that we do not assume that (2.9) holds for $r = +\infty$, which enables the convergence result to apply in the cases considered in the numerical section.

We then consider a sequence $(v_{\mathcal{T}^{(m)}})_{m\geq 1}$ of solutions with respect to the sequence $\left(\mathcal{T}^{(m)}\right)_{m\geq 1}$ and we study the convergence of this sequence to a solution of the continuous problem in the sense of Definition 1.1. Thus the second main result of this section is that this sequence converges, up to a subsequence, to a weak solution of the continuous problem in the sense of Definition 1.1, as stated by the following theorem.

Theorem 2.4. *Let $\left(\mathcal{T}^{(m)}\right)_{m\geq 1}$ be a sequence of meshes of the computational domain Ω in the sense specified in Section 2.2 such that (2.7) and (2.8) are satisfied and let $(\Lambda_{\mathcal{T}^{(m)}})_{m\geq 1}$ be such that (2.4), (2.5) and (2.9) hold. For any $m \geq 1$, let $v_{\mathcal{T}^{(m)}}$ be an arbitrary numerical solution in the sense of Definition 2.2 in the case where $\mathcal{T} = \mathcal{T}^{(m)}$ (the existence of $v_{\mathcal{T}^{(m)}}$ is given by Thm. 2.3). Then, there exists $\overline{u} \in \mathcal{S}_d(\Omega)$, weak solution of problems (1.1) and (1.2) in the sense of Definition 1.1, such that, up to the extraction of a subsequence, $(\psi(v_{\mathcal{T}^{(m)}}))_{m\geq 1}$ converges to $\overline{u}$ weakly in $W_0^{1,r}(\Omega)$ for all $r \in \left(1, \frac{d}{d-1}\right)$, strongly in $L^q(\Omega)$ for all $q \in [1,+\infty)$ if $d = 2$ and for all $q \in [1, d/(d-2))$ if $d > 2$, and almost everywhere in Ω.*

Moreover, for all $k > 0$, $T_k\overline{u} \in H_0^1(\Omega)$ holds.

In the case $d = 2$, the whole sequence converges to the unique solution of the problem.

Remark 2.5. Note also that $\overline{u} = \psi(v)$ where v is the limit of the subsequence $(v_{\mathcal{T}^{(m)}})_{m\geq 1}$ in $L^q(\Omega)$ for any $q \in [1,+\infty)$ if $d = 2$ or $q \in [1, \frac{2d}{d-2})$ if $d > 2$.

The remaining of this section is dedicated to the proof of Theorems 2.3 and 2.4.

2.4. Estimates

As done in the introduction, we denote in the whole paper by $\|\cdot\|_r$ the norm in $L^r(\Omega)$ or in $L^r(\Omega)^d$ for any $r \in [1,+\infty]$. Following the arguments presented in the introduction, we define the following function

$$\beta : \left| \begin{array}{l} (-1,1) \longrightarrow \mathbb{R} \\ s \longmapsto \int_0^s \sqrt{\psi'(t)}\,\mathrm{d}t. \end{array} \right. \tag{2.10}$$

In Figure 1, we propose a graphical representation of functions ψ, ψ' and β.

The following lemma is used in the course of the following computations.

Lemma 2.6. *For any $\varepsilon \in (0,1)$, there exists strictly positive reals a_ε, b_ε and c_ε only depending on ε, such that*

$$\forall s \in (-1,1),\ \sqrt{\psi'(s)} \leq a_\varepsilon |\beta(s)|^{(1+\varepsilon)/(1-\varepsilon)} + b_\varepsilon \text{ and } \exp\frac{(1-\varepsilon)|s|}{2(1-|s|)} \leq c_\varepsilon|\beta(s)| + 1. \tag{2.11}$$

Proof. We have, for all $\varepsilon \in (0,1)$ and $s \in (-1,1)$,

$$\sqrt{\psi'(s)} = \frac{1}{1-|s|}\exp\frac{|s|}{2(1-|s|)} \leq C_\varepsilon \exp\frac{(1+\varepsilon)|s|}{2(1-|s|)}, \tag{2.12}$$

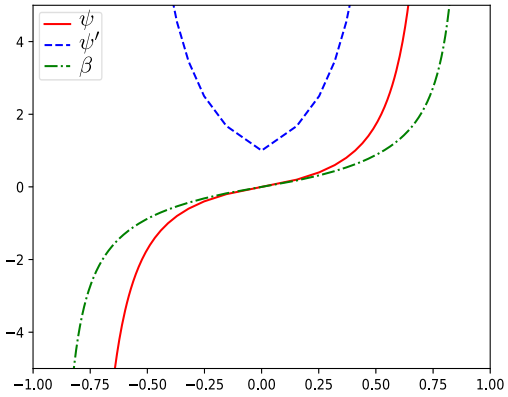

FIGURE 1. Functions ψ, ψ' and β.

where

$$C_\varepsilon = \max_{t\in(-1,1)} \frac{1}{1-|t|} \exp\Big(-\frac{\varepsilon|t|}{2(1-|t|)}\Big) = \frac{2}{\varepsilon}\exp\Big(-1+\frac{\varepsilon}{2}\Big).$$

Besides, we have

$$\begin{aligned}\exp\frac{(1-\varepsilon)|s|}{2(1-|s|)} &= \int_0^{|s|} \frac{1-\varepsilon}{2(1-|t|)^2}\exp\frac{(1-\varepsilon)|t|}{2(1-|t|)}\,\mathrm{d}t + 1\\ &\le \frac{(1-\varepsilon)}{2}C_\varepsilon\int_0^{|s|}\frac{1}{1-|t|}\exp\frac{|t|}{2(1-|t|)}\,\mathrm{d}t + 1 = \frac{(1-\varepsilon)}{2}C_\varepsilon|\beta(s)| + 1. \end{aligned} \tag{2.13}$$

Consequently we obtain that

$$\exp\frac{(1+\varepsilon)|s|}{2(1-|s|)} \le \left(\frac{(1-\varepsilon)}{2}C_\varepsilon|\beta(s)|+1\right)^{(1+\varepsilon)/(1-\varepsilon)}.$$

Owing to (2.12), this gives

$$\forall s\in(-1,1),\ \sqrt{\psi'(s)} \le C_\varepsilon\left(\frac{(1-\varepsilon)}{2}C_\varepsilon|\beta(s)|+1\right)^{(1+\varepsilon)/(1-\varepsilon)},$$

which provides the conclusion of the lemma owing to a convexity argument. □

We have the following estimate.

Lemma 2.7. *Let $v_\mathcal{T}\in\mathcal{V}_\mathcal{T}(\Omega)$ be such that*

$$\max_{x\in\overline{\Omega}}|v_\mathcal{T}(x)|<1 \ \text{ and } \ \int_\Omega \psi'(v_\mathcal{T})\Lambda_\mathcal{T}\nabla v_\mathcal{T}\cdot\nabla v_\mathcal{T}\,\mathrm{d}x \le \int_\Omega v_\mathcal{T}\,\mathrm{d}f. \tag{2.14}$$

Then there exists C_1 depending only on $\underline{\lambda}$ and $\|f\|_{M(\Omega)}$ such that

$$\|\nabla\beta(v_\mathcal{T})\|_2 \le C_1. \tag{2.15}$$

Proof. On one hand, using the fact that $\max_{x\in\overline{\Omega}}|v_\mathcal{T}(x)|<1$ we have

$$\int_\Omega v_\mathcal{T}\,\mathrm{d}f \le \|f\|_{M(\Omega)},$$

and on the other hand, accounting for (2.5), we have

$$\underline{\lambda}\int_\Omega \psi'(v_\mathcal{T})|\nabla v_\mathcal{T}|^2\,\mathrm{d}x \le \int_\Omega \psi'(v_\mathcal{T})\Lambda_\mathcal{T}\nabla v_\mathcal{T}\cdot\nabla v_\mathcal{T}\,\mathrm{d}x.$$

Hence we get

$$\underline{\lambda}\|\nabla\beta(v_\mathcal{T})\|_2^2 \le \|f\|_{M(\Omega)}.$$

□

The next lemma provides a lower bound on $1-|v_\mathcal{T}|$ for any function $v_\mathcal{T}$ such that the estimate (2.14) holds.

Lemma 2.8. *Let $v_\mathcal{T}\in\mathcal{V}_\mathcal{T}(\Omega)$ be such that (2.14) holds. Then there exists $C_2\in(0,1)$ depending only on $\underline{\lambda}, \|f\|_{M(\Omega)}, d$, the measure $|\Omega|$ of Ω and on $\mathcal{T}$ such that*

$$\max_{x\in\overline{\Omega}}|v_\mathcal{T}(x)|\le C_2.$$

Proof. Owing to $\beta(v_\mathcal{T})\in H_0^1(\Omega)$ and Lemma A.1 with $q=1$, we can write

$$\|\beta(v_\mathcal{T})\|_1 \le C_{\mathrm{sob}}^{(2,1)}\|\nabla\beta(v_\mathcal{T})\|_2.$$

By virtue of (2.11) with $\varepsilon=1/2$, together with Lemma 2.7, we obtain,

$$\int_\Omega \exp\left(\frac{1}{4}\frac{|v_\mathcal{T}|}{1-|v_\mathcal{T}|}\right)\mathrm{d}x \le c_{1/2}C_{\mathrm{sob}}^{(2,1)}C_1+|\Omega|.$$

Let $n\in\mathbb{N}$ which will be chosen later, and let us define $M_n = \min_{s\in(-1,1)}(1-|s|)^n\exp\left(\frac{1}{4}\frac{|s|}{1-|s|}\right) = \frac{1}{(4n)^n}\exp\left(n-\frac{1}{4}\right)$. Note that we have $M_n>0$ and

$$\int_\Omega \frac{M_n}{(1-|v_\mathcal{T}|)^n}\,\mathrm{d}x \le c_{1/2}C_{\mathrm{sob}}^{(2,1)}C_1+|\Omega|.$$

Since $|v_\mathcal{T}(z_i)|<1$ for any $i\in\mathcal{N}$, and using the fact that

$$v_\mathcal{T}=\sum_{i\in\mathcal{N}} v_\mathcal{T}(z_i)\varphi_i,$$

we get that, for *a.e.* $x\in\Omega$,

$$|\nabla v_\mathcal{T}(x)|\le \operatorname*{esssup}_{y\in\Omega}\sum_{i\in\mathcal{N}}|\nabla\varphi_i(y)|.$$

Let $i_0\in\mathcal{N}$ be such that $1-|v_\mathcal{T}(z_{i_0})| = \min_{i\in\mathcal{N}}(1-|v_\mathcal{T}(z_i)|)$. Then denoting $G_\mathcal{T}=\operatorname{esssup}_{y\in\Omega}\sum_{i\in\mathcal{N}}|\nabla\varphi_i(y)|$ we have, for any $x\in\Omega$,

$$1-|v_\mathcal{T}(x)|\le 1-|v_\mathcal{T}(z_{i_0})|+G_\mathcal{T}|x-z_{i_0}|,$$

which implies

$$\int_\Omega \frac{M_n}{(1-|v_\mathcal{T}|)^n}\,\mathrm{d}x \ge \int_\Omega \frac{M_n}{(1-|v_\mathcal{T}(z_{i_0})|+G_\mathcal{T}|x-z_{i_0}|)^n}\,\mathrm{d}x.$$

Let $r_\mathcal{T}>0$ be such that $B(z_i,r_\mathcal{T})\subset\Omega$ for all $i\in\mathcal{N}$, and set $R_\mathcal{T}=(1-|v_\mathcal{T}(z_{i_0})|)r_\mathcal{T}\le r_\mathcal{T}$. Then

$$\int_\Omega \frac{M_n}{(1-|v_\mathcal{T}(z_{i_0})|+G_\mathcal{T}|x-z_{i_0}|)^n}\,\mathrm{d}x \ge \int_{B(z_{i_0},R_\mathcal{T})}\frac{M_n}{(1-|v_\mathcal{T}(z_{i_0})|+G_\mathcal{T}|x-z_{i_0}|)^n}\,\mathrm{d}x$$

$$\begin{aligned} &\geq \frac{\gamma_d R_{\mathcal{T}}^d}{(1+G_{\mathcal{T}}\ r_{\mathcal{T}})^n} \frac{M_n}{(1-|v_{\mathcal{T}}(z_{i_0})|)^n} \\ &= \frac{\gamma_d r_{\mathcal{T}}^d}{(1+G_{\mathcal{T}}\ r_{\mathcal{T}})^n} \frac{M_n}{(1-|v_{\mathcal{T}}(z_{i_0})|)^{n-d}}, \end{aligned} \tag{2.16}$$

where $\gamma_d > 0$ only depending on d is such that $|B(z_{i_0}, R_{\mathcal{T}})| = \gamma_d {R_{\mathcal{T}}}^d$ and

$$\forall x \in B(z_{i_0}, R_{\mathcal{T}}), \quad 1 - |v_{\mathcal{T}}(z_{i_0})| + G_{\mathcal{T}}|x - z_{i_0}| \leq (1 - |v_{\mathcal{T}}(z_{i_0})|)(1 + G_{\mathcal{T}}\ r_{\mathcal{T}}).$$

We then obtain

$$\frac{\gamma_d r_{\mathcal{T}}^d}{(1+G_{\mathcal{T}}\ r_{\mathcal{T}})^n} \frac{M_n}{(1-|v_{\mathcal{T}}(z_{i_0})|)^{n-d}} \leq C_{\mathrm{sob}}^{(2,1)} C_1 + |\Omega|,$$

which gives

$$\frac{\gamma_d r_{\mathcal{T}}^d}{(1+G_{\mathcal{T}}\ r_{\mathcal{T}})^n} \frac{M_n}{C_{\mathrm{sob}}^{(2,1)} C_1 + |\Omega|} \leq (1 - |v_{\mathcal{T}}(z_{i_0})|)^{n-d}.$$

Taking $n = d + 1$ in the previous inequality gives

$$\frac{\gamma_d r_{\mathcal{T}}^d}{(1+G_{\mathcal{T}}\ r_{\mathcal{T}})^{d+1}} \frac{M_{d+1}}{C_{\mathrm{sob}}^{(2,1)} C_1 + |\Omega|} \leq 1 - |v_{\mathcal{T}}(z_{i_0})|,$$

which concludes the proof. □

Lemma 2.9. *Let $r \in [1, d/(d-1))$. Let $v_{\mathcal{T}} \in \mathcal{V}_{\mathcal{T}}(\Omega)$ be such that (2.14) holds. Then, there exists C_3 only depending on r, d, $\underline{\lambda}$ and $\|f\|_{M(\Omega)}$ such that*

$$\|\nabla \psi(v_{\mathcal{T}})\|_r \leq C_3.$$

Proof. We have

$$\|\nabla \psi(v_{\mathcal{T}})\|_r^r = \int_\Omega |\psi'(v_{\mathcal{T}})|^r |\nabla v_{\mathcal{T}}|^r \,\mathrm{d}x = \int_\Omega |\psi'(v_{\mathcal{T}})|^{r/2} |\nabla \beta(v_{\mathcal{T}})|^r \,\mathrm{d}x.$$

Thanks to Hölder's inequality, letting $a = 2/r$, $b = 2/(2-r)$, we get, owing to Lemma 2.7,

$$\|\nabla \psi(v_{\mathcal{T}})\|_r^r \leq \|\nabla \beta(v_{\mathcal{T}})\|_2^r \|\psi'(v_{\mathcal{T}})\|_{r/(2-r)}^{r/2} \leq (C_1)^r \|\psi'(v_{\mathcal{T}})\|_{r/(2-r)}^{r/2}.$$

Owing to (2.11), we can write, for any $\varepsilon \in (0,1)$, and applying Hölder's inequality,

$$\|\psi'(v_{\mathcal{T}})\|_{r/(2-r)}^{r/(2-r)} \leq \int_\Omega \Big(a_\varepsilon |\beta(v_{\mathcal{T}})|^{(1+\varepsilon)/(1-\varepsilon)} + b_\varepsilon \Big)^{2r/(2-r)} \,\mathrm{d}x \leq a_{\varepsilon,r} \|\beta(v_{\mathcal{T}})\|_q^q + b_{\varepsilon,r},$$

with $a_{\varepsilon,r} > 0$ and $b_{\varepsilon,r} > 0$ only depending on ε and r, and

$$q = \frac{2r(1+\varepsilon)}{(2-r)(1-\varepsilon)}.$$

Let us now choose ε in order that, for the above value of q, we can apply the Sobolev inequality provided by Lemma A.1 to the function $\beta(v_{\mathcal{T}}) \in H_0^1(\Omega)$ and Lemma 2.7, which gives

$$\|\beta(v_{\mathcal{T}})\|_q \leq C_{\mathrm{sob}}^{(2,q)} \|\nabla \beta(v_{\mathcal{T}})\|_2 \leq C_{\mathrm{sob}}^{(2,q)} C_1. \tag{2.17}$$

– In the case $d = 2$, we can choose $\varepsilon = \frac{1}{2}$.

– In the case $d > 2$, let us select $\varepsilon \in (0,1)$ such that

$$q = \frac{2r}{2-r}\frac{1+\varepsilon}{1-\varepsilon} \leq \frac{2d}{d-2}. \tag{2.18}$$

Since $r \in [1, d/(d-1))$ implies $(2-r)/r \in ((d-2)/d, 1]$, the quantity a_r such that $a_r = \frac{1}{2}\Big(1 + \frac{d}{d-2}\frac{2-r}{r}\Big)$ is such that $1 < a_r \leq \frac{d}{d-2}\frac{2-r}{r}$. It suffices to choose $\varepsilon = \frac{a_r - 1}{a_r + 1}$ for obtaining both $\varepsilon \in (0,1)$ and (2.18).

For this value of ε, which only depends on d and r, we get the conclusion of the lemma. □

2.5. Existence of a solution to Scheme (2.6)

The purpose of this section is to prove Theorem 2.3, which states the existence of a solution to the numerical scheme in the sense of Definition 2.2, by applying the topological degree method [12].

Proof of Theorem 2.3. Let us define the continuous function

$$\mathcal{L}: \left| \begin{array}{l} \mathbb{R}^{\mathcal{N}_{\rm int}} \longrightarrow \mathcal{V}_{\mathcal{T}}(\Omega) \\ (u_i)_{i\in\mathcal{N}_{\rm int}} \longmapsto \displaystyle\sum_{i\in\mathcal{N}_{\rm int}} \psi^{-1}(u_i)\varphi_i, \end{array} \right. \tag{2.19}$$

and the function

$$\mathcal{F}: \left| \begin{array}{l} \mathbb{R}^{\mathcal{N}_{\rm int}} \times [0,1] \longrightarrow \mathbb{R}^{\mathcal{N}_{\rm int}} \\ (u,\mu) \longmapsto \mathcal{F}(u,\mu) = (\mathcal{F}_i(u,\mu))_{i\in\mathcal{N}_{\rm int}}, \end{array} \right. \tag{2.20}$$

where for any $u = (u_j)_{j\in\mathcal{N}_{\rm int}} \in \mathbb{R}^{\mathcal{N}_{\rm int}}$, $\mu \in [0,1]$ and $i \in \mathcal{N}_{\rm int}$, the quantity $\mathcal{F}_i(u,\mu)$ is defined by

$$\mathcal{F}_i(u,\mu) := \mu\bigg(\int_\Omega \psi'(\mathcal{L}(u))\Lambda_{\mathcal{T}}\nabla\mathcal{L}(u)\cdot\nabla\varphi_i\,\mathrm{d}x - \int_\Omega \varphi_i\,\mathrm{d}f\bigg) + (1-\mu)u_i.$$

This mapping is well defined and continuous, since, for any $u = (u_i)_{i\in\mathcal{N}_{\rm int}} \in \mathbb{R}^{\mathcal{N}_{\rm int}}$, we have $\max_{x\in\overline{\Omega}} |\mathcal{L}(u)(x)| \leq \max_{i\in\mathcal{N}_{\rm int}} |\psi^{-1}(u_i)| < 1$ and $\mathcal{L}$ is continuous. We also notice that the equation $\mathcal{F}(u,1) = 0$ gives

$$\forall i \in \mathcal{N}_{\rm int},\ \int_\Omega \psi'(\mathcal{L}(u))\Lambda_{\mathcal{T}}\nabla\mathcal{L}(u)\cdot\nabla\varphi_i\,\mathrm{d}x = \int_\Omega \varphi_i\,\mathrm{d}f.$$

In particular we obtain

$$\int_\Omega \psi'(\mathcal{L}(u))\Lambda_{\mathcal{T}}\nabla\mathcal{L}(u)\cdot\nabla w\,\mathrm{d}x = \int_\Omega w\,\mathrm{d}f \text{ for any } w \in \mathcal{V}_{\mathcal{T}}(\Omega),$$

which means that $\mathcal{L}(u)$ is a solution of the numerical scheme (2.6). Let $\mu \in (0,1]$ and let $u = (u_i)_{i\in\mathcal{N}_{\rm int}} \in \mathbb{R}^{\mathcal{N}_{\rm int}}$ be such that $\mathcal{F}(u,\mu) = 0$. We have, for any $i \in \mathcal{N}_{\rm int}$,

$$\mu\int_\Omega \psi'(\mathcal{L}(u))\Lambda_{\mathcal{T}}\nabla\mathcal{L}(u)\cdot\nabla\varphi_i\,\mathrm{d}x + (1-\mu)u_i = \mu\int_\Omega \varphi_i\,\mathrm{d}f.$$

Multiplying the previous identity by $\psi^{-1}(u_i)$ and summing over the internal nodes leads to

$$\mu\int_\Omega \psi'(\mathcal{L}(u))\Lambda_{\mathcal{T}}\nabla\mathcal{L}(u)\cdot\nabla\mathcal{L}(u)\,\mathrm{d}x + (1-\mu)\sum_{i\in\mathcal{N}} u_i\psi^{-1}(u_i) = \mu\int_\Omega \mathcal{L}(u)\,\mathrm{d}f.$$

Dividing by μ and using $\frac{1-\mu}{\mu}\sum_{i\in\mathcal{N}_{\rm int}} u_i\psi^{-1}(u_i)\geq 0$, we get that

$$\int_\Omega \psi'(\mathcal{L}(u))\Lambda_{\mathcal{T}}\nabla\mathcal{L}(u)\cdot\nabla\mathcal{L}(u)\,\mathrm{d}x \leq \int_\Omega \mathcal{L}(u)\,\mathrm{d}f,$$

which means that (2.14) holds for $\mathcal{L}(u)$. Hence, from Lemma 2.8, we then obtain

$$\max_{x\in\overline{\Omega}}|\mathcal{L}(u)(x)|\leq C_2.$$

We then obtain that $|\psi^{-1}(u_i)|\leq C_2$ for all $i\in\mathcal{N}_{\rm int}$, and therefore that

$$|u_i|\leq\psi(C_2),\ \forall i\in\mathcal{N}_{\rm int}.$$

Define the relatively compact open set

$$\mathcal{U}=\left\{u=(u_i)_{i\in\mathcal{N}}\in\mathbb{R}^{\mathcal{N}_{\rm int}}\text{ such that }|u_i|<\psi(C_2)+1\right\}.$$

For $\mu=0$, the linear equation $\mathcal{F}(u,0)=0$ has the unique solution $u=0$. The topological degree corresponding to $\mathcal{F}(u,0)$ and $\mathcal{U}$ is therefore equal to 1 since $u=0$ belongs to $\mathcal{U}$. Hence for any $\mu\in[0,1]$ any solution to $\mathcal{F}(u,\mu)=0$ necessarily belongs to $\mathcal{U}$. Therefore, owing to the invariance of the topological degree by homotopy, there exists at least one $u\in\mathbb{R}^{\mathcal{N}_{\rm int}}$ (non necessarily unique) such that $\mathcal{F}(u,1)=0$, which means that $\mathcal{L}(u)$ is a solution to Scheme (2.6) in the sense of Definition 2.2. □

2.6. Convergence of the nonlinear finite element scheme

The goal of this section is the proof of Theorem 2.4, which uses the following lemma.

Lemma 2.10. *Let Ω be an open bounded subset of $\mathbb{R}^d$ with $d\geq 2$. Let $\left(u^{(m)}\right)_{m\geq 0}\in\mathcal{S}_d(\Omega)^{\mathbb{N}}$ such that there exists $(C_r)_{r\geq 0}\in(\mathbb{R}_+)^{\mathbb{N}}$ satisfying, for any $r\in\left(1,\frac{d}{d-1}\right)$,*

$$\left\|\nabla u^{(m)}\right\|_r\leq C_r,\ \text{for any } m\geq 0.$$

Then there exists $\overline{u}\in\mathcal{S}_d(\Omega)$ such that, up to a subsequence, $\left(u^{(m)}\right)_{m\geq 0}$ weakly converges to $\overline{u}$ in all $W_0^{1,r}(\Omega)$ for all $r\in(1,d/(d-1))$, strongly in $L^q(\Omega)$ for all $q\in[1,+\infty)$ if $d=2$ and for all $q\in[1,d/(d-2))$ if $d\geq 3$, and almost everywhere in Ω.

Proof. Let us first consider a given $r_0\in\left(1,\frac{d}{d-1}\right)$. Let us select $\overline{u}\in W_0^{1,r_0}(\Omega)$ and an infinite set $S_1\subset\mathbb{N}$, such that $\left(u^{(m)}\right)_{m\in S_1}$ weakly converges to $\overline{u}$ in $W_0^{1,r_0}(\Omega)$. Let $r\in(1,d/(d-1))$ be given. There exists $\overline{u}'\in W_0^{1,r}(\Omega)$ and an infinite set $S_2\subset S_1$ such that $\left(u^{(m)}\right)_{m\in S_2}$ weakly converges to $\overline{u}'$ in $W_0^{1,r}(\Omega)$. Then $\overline{u}=\overline{u}'$ in $W_0^{1,\min(r,r_0)}(\Omega)$, which implies, by uniqueness of the limit, that we can take $S_2=S_1$, and that $\overline{u}\in W_0^{1,\max(r,r_0)}(\Omega)$. Since this holds for any $r\in(1,d/(d-1))$, the lemma is proved using Sobolev inequalities. □

We can now prove the weak convergence of a discrete solution to a solution of the continuous problem.

Proof of Theorem 2.4. For all $m\geq 1$, we let $w=v_{\mathcal{T}^{(m)}}$ in (2.6) (recall that this would not be possible if we were considering a simple linear finite element approximation of this problem, as we do for comparison purposes in the numerical section). Hence (2.14) holds (it is in this case an equality instead of an inequality), and we can apply Lemma 2.9. Therefore the sequence $(\psi(v_{\mathcal{T}^{(m)}}))_{m\geq 1}$ is bounded in $W_0^{1,r}(\Omega)$ for any $r\in[1,\frac{d}{d-1})$. Applying Lemma 2.10, we get that there exists $\overline{u}\in\mathcal{S}_d(\Omega)$ such that, up to the extraction of a subsequence, $(\psi(v_{\mathcal{T}^{(m)}}))_{m\geq 1}$ weakly converges to $\overline{u}$ in all $W_0^{1,r}(\Omega)$ for all $r\in\left(1,\frac{d}{d-1}\right)$.

For all $m \in \mathbb{N}$, we define the following interpolation operator

$$\mathcal{P}_{\mathcal{T}^{(m)}} : \left| \begin{array}{l} C_c^\infty(\Omega) \longrightarrow \mathcal{V}_{\mathcal{T}^{(m)}}(\Omega) \\ v \longmapsto \displaystyle\sum_{i \in \mathcal{N}^{(m)}} v(z_i^{(m)}) \varphi_i^{(m)}. \end{array} \right. \tag{2.21}$$

This operator satisfies the following approximation properties (see [16]): for any $\varphi \in C_c^\infty(\Omega)$,

$$\|\varphi - \mathcal{P}_{\mathcal{T}^{(m)}}(\varphi)\|_\infty + h_{\mathcal{T}^{(m)}} \|\nabla\varphi - \nabla\mathcal{P}_{\mathcal{T}^{(m)}}(\varphi)\|_\infty \le C_{\rm inter} h_{\mathcal{T}^{(m)}}^2 \sum_{k=1}^d \sum_{\ell=1}^d \left\| \frac{\partial^2 \varphi}{\partial x_k \partial x_\ell} \right\|_\infty, \tag{2.22}$$

where $C_{\rm inter}$ is increasingly depending on $\theta_{\mathcal{T}^{(m)}}$. For a given $\varphi \in C_c^\infty(\Omega)$, we let $w = \mathcal{P}_{\mathcal{T}^{(m)}}(\varphi)$ in (2.6). We get for any $m \ge 1$,

$$\int_\Omega \Lambda_{\mathcal{T}^{(m)}} \nabla\psi(v_{\mathcal{T}^{(m)}}) \cdot \nabla\mathcal{P}_{\mathcal{T}^{(m)}}(\varphi)\,{\rm d}x = \int_\Omega \mathcal{P}_{\mathcal{T}^{(m)}}(\varphi)\,{\rm d}f.$$

Owing to (2.22) and (2.8), we get that the sequence $(\mathcal{P}_{\mathcal{T}^{(m)}}(\varphi))_{m\ge1}$ converges to φ in $L^\infty(\Omega)$ and that the sequence $(\Lambda_{\mathcal{T}^{(m)}}\nabla\mathcal{P}_{\mathcal{T}^{(m)}}(\varphi))_{m\ge1}$ converges to $\Lambda\nabla\varphi$ in $L^r(\Omega)$ for all $r \in [1,+\infty)$ owing to (2.9) and to the convergence of $(\nabla\mathcal{P}_{\mathcal{T}^{(m)}}(\varphi))_{m\ge1}$ to $\nabla\varphi$ in $L^\infty(\Omega)$. Since the sequence $(\psi(v_{\mathcal{T}^{(m)}}))_{m\ge1}$ weakly converges to $\overline{u}$ in all $W_0^{1,r}(\Omega)$ for all $r \in (1, d/(d-1))$ we then obtain that

$$\lim_{m\to\infty} \int_\Omega \Lambda_{\mathcal{T}^{(m)}} \nabla\psi(v_{\mathcal{T}^{(m)}}) \cdot \nabla\mathcal{P}_{\mathcal{T}^{(m)}}(\varphi)\,{\rm d}x = \int_\Omega \Lambda\nabla\overline{u}\cdot\nabla\varphi\,{\rm d}x.$$

The continuity property of $f \in M(\Omega)$ implies

$$\lim_{m\to\infty} \int_\Omega \mathcal{P}_{\mathcal{T}^{(m)}}(\varphi)\,{\rm d}f = \int_\Omega \varphi\,{\rm d}f.$$

Consequently, equation (1.5) holds for any $\varphi \in C_c^\infty(\Omega)$. By a density argument we obtain that (1.5) holds for any $w \in \mathcal{T}_d(\Omega)$, which gives that $\overline{u}$ is a weak solution of problems (1.1) and (1.2) in the sense of Definition 1.1.

Let us now prove that, for all $k > 0$, $T_k\overline{u} \in H_0^1(\Omega)$. Using that $T_k'(s) = 0$ for $|s| > k$ and $T_k'(s) = 1$ for $|s| \le k$, we have that

$$\begin{aligned} \int_\Omega |\nabla T_k u_{\mathcal{T}^{(m)}}|^2\,{\rm d}x &= \int_\Omega |\nabla T_k\psi(v_{\mathcal{T}^{(m)}})|^2\,{\rm d}x = \int_\Omega \psi'(v_{\mathcal{T}^{(m)}})^2 T_k'(\psi(v_{\mathcal{T}^{(m)}}))^2 |\nabla v_{\mathcal{T}^{(m)}}|^2\,{\rm d}x \\ &\le \frac{\psi'\big(\psi^{-1}(k)\big)}{\underline{\lambda}} \int_\Omega \Lambda_{\mathcal{T}^{(m)}} \psi'(v_{\mathcal{T}^{(m)}}) \nabla v_{\mathcal{T}^{(m)}} \cdot \nabla v_{\mathcal{T}^{(m)}}\,{\rm d}x \le \frac{\psi'\big(\psi^{-1}(k)\big)}{\underline{\lambda}} \|f\|_{M(\Omega)}. \end{aligned} \tag{2.23}$$

Using the above inequality, selecting a convenient subsequence, we have

$$\int_\Omega |\nabla T_k\overline{u}|^2\,{\rm d}x \le \liminf_{m\to\infty} \int_\Omega |\nabla T_k u_{\mathcal{T}^{(m)}}|^2\,{\rm d}x,$$

which leads to $T_k\overline{u} \in H_0^1(\Omega)$. □

3. Study of the nonlinear CVFE scheme

3.1. Motivation and organisation

As already noticed, for some diffusion fields Λ satisfying Hypothesis (1.3b), there exist non-zero weak solutions to problems (1.1) and (1.2) in the sense of Definition 1.1 for $f = 0$. One such diffusion field is precisely considered in the numerical examples presented in Section 4. But, among these solutions, there is one and only one which

is in some sense the limit of more regular problems, as proved in [3]. This solution must satisfy a regularity criterion which is the basis of the notion of entropy weak solution, in the case where $f \in L^1(\Omega)$. We give in Definition 3.1 below this entropy weak solution sense, formulated in the particular case of bounded domains and linear uniformly elliptic operators.

Definition 3.1. Entropy solution to the linear elliptic problem with right-hand side in $L^1(\Omega)$.

Let $\mathcal{S}_d(\Omega)$ be the set of functions defined in (1.4), and, for all $k > 0$, let $T_k : \mathbb{R} \to \mathbb{R}$ be the truncation function defined by $s \mapsto \min(|s|, k)\mathrm{sign}(s)$. We assume that $f \in L^1(\Omega)$.

An entropy solution of problems (1.1) and (1.2) in the sense of [3] is a function $\overline{u} \in \mathcal{S}_d(\Omega)$ satisfying that

(1) for any $k > 0$, $T_k(\overline{u}) \in H_0^1(\Omega)$,
(2) for any $k > 0$ and for any $\phi \in C_c^\infty(\Omega)$,

$$\int_{|\overline{u}-\phi|\le k} \Lambda\nabla\overline{u}\cdot\nabla(\overline{u}-\phi)\,\mathrm{d}x \le \int_\Omega T_k(\overline{u}-\phi)f(x)\,\mathrm{d}x. \tag{3.1}$$

Remark 3.2. It is proved in [3] that it is equivalent to replace (3.1) and the truncations T_k by

$$\int_\Omega \Lambda\nabla\overline{u}\cdot\nabla T(\overline{u}-\phi)\,\mathrm{d}x \le \int_\Omega T(\overline{u}-\phi)f(x)\,\mathrm{d}x, \tag{3.2}$$

for any $\phi \in C_c^\infty(\Omega)$ and for any $T \in \mathcal{F}$, where $\mathcal{F}$ is defined as the set of all functions $T \in C^2(\mathbb{R}) \cap L^\infty(\mathbb{R})$ such that

$$T(0) = 0, \quad T' \ge 0, \quad T'(s) = 0 \text{ for } s \text{ large enough};$$

$$T(-s) = -T(s), \quad T''(s) \le 0 \ \text{ for } s \ge 0.$$

It is also proved in [3] that the entropy solution exists, is unique and is a weak solution in the sense of Definition 1.1.

Remark 3.3. In fact, Andrea Dall'Aglio proved in 1996 that the inequality (3.1) can be replaced by an equality [11]. This result is stated and briefly proved in Lemma A.6 in the appendix, as well as the fact that (3.2) can also be replaced by an equality.

The above entropy solution sense is reviewed in Prignet [23], as well as different mathematical senses for a solution to this problem. It is proved to be equivalent to the renormalised sense introduced in [8,10]. See [3,4,25] for more detailed definitions and properties; one can also refer to [13,14] for some extensions to the case where the problem is not coercive.

Recall that, in [8], the authors prove the convergence of the approximate solution obtained by the linear finite element approximation (1.6) to the entropy (or renormalised) solution, but only under strong restrictions on the meshes and Λ-fields that can be considered. These restrictions are not requested for the nonlinear finite element scheme studied in Section 2 of this work, which is shown to converge to a weak solution of the problem on any simplicial grid, with any diffusion field. Nevertheless, the convergence of this nonlinear finite element scheme to the entropy solution of the problem could not be proved, mainly because no inequality can be obtained by introducing nonlinear functions of the primary unknown as test functions.

We therefore propose in Section 3 of this paper a nonlinear scheme which holds for any simplicial grid and for any diffusion field Λ satisfying Hypothesis (1.3b). We construct this scheme in such a way that we can prove its convergence to the entropy weak solution of the problem, owing to three ideas.

(1) The first one is motivated in Section 2.1 above: it relies on writing $\overline{u} = \psi(v)$, using the function ψ defined by (2.2). Then, as in Section 2, an estimate is deduced when one takes the approximation of the bounded function $\overline{v}$ as test function in the numerical scheme.

(2) The second one is the use of a control volume-finite element scheme (called for short CVFE in this work), which allows to simultaneously adopt the finite element point of view, for computing the coefficients of a rigidity matrix, and the finite volume point of view, for the computation of some nonlinear expressions with respect to the unknown.
Then, following [2,6,7] and other works by the same authors, the evaluation of ψ' is upstream weighted with respect to the sign of the so-called "transmissibility" between control volumes (this quantity, which only depends on the mesh and on Λ, is defined by (3.4)). Owing to this technique, it becomes possible to derive estimates with using test functions under the form of nonlinear functions of the primary unknown. For this purpose, these nonlinear functions must satisfy some specific properties (see (A.3)), which is restricting the range of the nonlinear functions which can be considered: for example, truncations T_k do not provide such monotony inequalities, see Remark 3.16.
As in Section 2, Sobolev inequalities (see Lem. A.1) play an important role for establishing these estimates; the application of these inequalities to the CVFE scheme is done through the equivalence property between continuous and discrete norms (see Lem. A.2). The discrete Sobolev inequalities as stated in [17] could also be used, up to the adaptation of the treatment of the homogeneous Dirichlet boundary conditions.

(3) We notice in Remark 3.11 that this upstream weighting yields a kind of weak p-Laplace stabilisation term with $p = 3$ (in analogy with the Godunov scheme for scalar hyperbolic equation, which provides a weak BV-inequality). Nevertheless, we could not conclude the convergence proof to the entropy solution only using this weak inequality, and we had to introduce in the scheme a nonlinear vanishing stabilisation term, based on a p-Laplace operator with $p > 3$. This term is not necessary for proving, up to a subsequence, the convergence of the scheme to a weak solution; but it is strongly used in the proof of the convergence to the entropy weak solution in the case where $f \in L^1(\Omega)$. One of the difficulties in the definition of this term is that it must be sufficiently large for bounding the expressions which must tend to zero, and sufficiently small for vanishing against regular test functions.

The organisation of Section 3 is similar to that of Section 2. In Section 3.2, we present the CVFE numerical scheme and give the main results (existence and convergence) concerning this scheme in Section 3.3. In Section 3.4, estimates on the discrete solution are established, including an inequality induced by the stabilisation term. These estimates allow, in Section 3.5, to derive the existence of a solution to the scheme, using a topological degree argument (similarly to what is done in Sect. 2). Section 3.6 is first devoted to the convergence proof of a subsequence of discrete solutions to a weak solution of the problem in the general case $f \in M(\Omega)$ (recall that the uniqueness of this solution holds only if $d = 2$). Then the convergence of the whole sequence to the entropy weak solution is proved in the case where $f \in L^1(\Omega)$. This last proof strongly relies on the presence of the p-Laplace stabilisation term. Let us observe that the complexity of these convergence proofs is largely greater than that of the convergence to a weak solution of the nonlinear finite element scheme studied in Section 2.

3.2. Definition of the scheme

The Control Volume Finite Element (CVFE) scheme is based on P^1-finite elements on a primal simplicial mesh, and on piecewise constant functions on a dual mesh (see Fig. 2 for an illustration in the case $d = 2$). Let us introduce the geometrical objects used in the definition of this scheme, which will be considered in the following to be all collected by the notation $\mathcal{T}$.

Primal mesh

We use the elements, nodes, P^1 basis functions defined in Section 2.2 which define the primal mesh of Ω.

Edges

We denote by $\mathcal{E}$ the set of all pairs $\{i, j\}$ such that there exists $K \in \mathcal{T}$ with $\{i, j\} \subset \mathcal{N}_K$. For any $\{i, j\} \in \mathcal{E}$, the length of the segment $[z_i, z_j]$ (called an edge of the mesh) is denoted by d_{ij}. For all $K \in \mathcal{T}$, we denote by $\mathcal{E}_K$

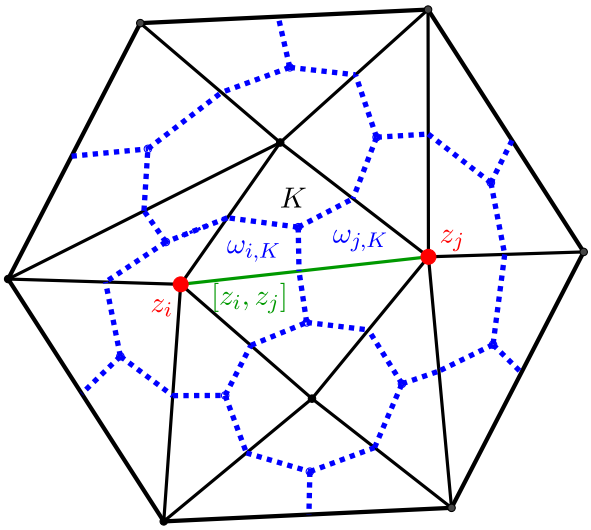

FIGURE 2. Triangular mesh $\mathcal{T}$ (solid line) and dual mesh $\mathcal{M}$ (dashed line).

the subset of $\mathcal{E}$ containing all the edges of K, and for any $\{i,j\} \in \mathcal{E}$, we denote by $\mathcal{T}_{ij}$ the set of all $K \in \mathcal{T}$ such that $\{i,j\} \in \mathcal{E}_K$. Accounting for the fact that any simplex has $d(d+1)/2$ edges, we define for any $\{i,j\} \in \mathcal{E}$,

$$m_{ij} = \frac{2}{d(d+1)} \sum_{K \in \mathcal{T}_{ij}} |K|.$$

Dual mesh

Once the primal triangular mesh has been built, we can define its dual barycentric mesh $\mathcal{M}$ as follows. To each $i \in \mathcal{N}$ and $K \in \mathcal{T}_i$, we define the set $\omega_{i,K}$ of all $x \in K$ such that $\varphi_i(x) > \varphi_j(x)$ for all $j \in \mathcal{N}_K \setminus \{i\}$ (we then have $\overline{K} = \bigcup_{i \in \mathcal{N}_K} \overline{\omega_{i,K}}$). Then we define $\omega_i = \bigcup_{K \in \mathcal{T}_i} \omega_{i,K}$ and $\mathcal{M} = \{\omega_i, i \in \mathcal{N}\}$.

Note that $\overline{\Omega} = \bigcup_{i \in \mathcal{N}} \overline{\omega}_i$ and $\omega_i \cap \omega_j = \emptyset$ for $i \neq j$. We refer to Figure 2 for an illustration of the primary and dual barycentric meshes in the 2D case. The Lebesgue measure of ω_i is denoted by m_i. The geometrical construction of ω_i ensures that

$$\int_\Omega \varphi_i(x)\,\mathrm{d}x = \int_{\omega_i} 1\,\mathrm{d}x =: m_i, \qquad \forall i \in \mathcal{N}.$$

We then denote by $\chi_{\omega_i} : \Omega \to \mathbb{R}$ the characteristic function of the subset ω_i for any $i \in \mathcal{N}$.

The space of all real families $(v_i)_{i \in \mathcal{N}}$ is classically denoted by $\mathbb{R}^{\mathcal{N}}$. Then we set

$$\mathbb{R}^{\mathcal{N}}_0 := \big\{(v_i)_{i \in \mathcal{N}} \in \mathbb{R}^{\mathcal{N}}, v_i = 0 \text{ for all } i \in \mathcal{N}_{\mathrm{ext}}\big\}.$$

For all $v \in \mathbb{R}^{\mathcal{N}}$ and for any continuous function $g : \mathbb{R} \to \mathbb{R}$, we denote by $g(v)$ the element of $\mathbb{R}^{\mathcal{N}}$ such that $(g(v))_i = g(v_i)$ for any $i \in \mathcal{N}$.

Given a family $v = (v_i)_{i \in \mathcal{N}} \in \mathbb{R}^{\mathcal{N}}$, we denote $\Pi_{\mathcal{T}} v \in C\big(\overline{\Omega}\big)$ and $\Pi_{\mathcal{M}} v \in L^1(\Omega)$ the functions defined by

$$\Pi_{\mathcal{T}} v = \sum_{i \in \mathcal{N}} v_i \varphi_i \quad \text{and} \quad \Pi_{\mathcal{M}} v = \sum_{i \in \mathcal{N}} v_i \chi_{\omega_i}.$$

Transmissibility coefficients

In order to define the CVFE scheme, we define the transmissibility coefficients (whose sign is not specified, contrary to the case of the finite volume-type schemes)

$$\Lambda_{ij}^K = -\int_K \Lambda \nabla\varphi_i \cdot \nabla\varphi_j \,\mathrm{d}x = \Lambda_{ji}^K, \quad \forall K \in \mathcal{T},\ \forall (i,j) \in \mathcal{N}^2, \tag{3.3}$$

and

$$\Lambda_{ij} = \Lambda_{ji} = -\int_\Omega \Lambda \nabla\varphi_i \cdot \nabla\varphi_j \,\mathrm{d}x = \sum_{K\in\mathcal{T}} \Lambda_{ij}^K, \quad \forall (i,j) \in \mathcal{N}^2. \tag{3.4}$$

Note that $\Lambda_{ij} = 0$ unless $\{i,j\} \in \mathcal{E}$. Moreover, since $\sum_{i\in\mathcal{N}_K} \nabla\varphi_i = 0$, we have that:

$$-\Lambda_{ii}^K = \sum_{j\in\mathcal{N}_K\setminus\{i\}} \Lambda_{ij}^K > 0. \tag{3.5}$$

As a consequence of (3.4) and (3.5), given v and w two elements of $\mathbb{R}_0^{\mathcal{N}}$, one has

$$\int_\Omega \Lambda \nabla\Pi_{\mathcal{T}} v \cdot \nabla\Pi_{\mathcal{T}} w \,\mathrm{d}x = \sum_{\{i,j\}\in\mathcal{E}} \Lambda_{ij}(v_i - v_j)(w_i - w_j) = \sum_{K\in\mathcal{T}} \sum_{\{i,j\}\in\mathcal{E}_K} \Lambda_{ij}^K (v_i - v_j)(w_i - w_j). \tag{3.6}$$

Upstream weighted CVFE scheme

We then introduce, for all $\{i,j\} \in \mathcal{E}$, the function

$$\Psi_{ij} : \left| \begin{array}{l} (-1,1)^2 \longrightarrow \mathbb{R} \\ (a,b) \longmapsto \begin{cases} \max_{s\in I(a,b)} \psi'(s) & \text{if } \Lambda_{ij} \geq 0 \\ \min_{s\in I(a,b)} \psi'(s) & \text{if } \Lambda_{ij} < 0, \end{cases} \end{array} \right. \tag{3.7}$$

where for any $a, b \in \mathbb{R}$, we denote by $I(a,b) = [\min(a,b), \max(a,b)]$.

We then consider the following approximation of problems (1.1) and (1.2).

Definition 3.4. Let $a_0 \in [0, +\infty)$ and $p \in (2, +\infty)$ be given. We say that v is a solution of the numerical scheme if there holds $v \in \mathbb{R}_0^{\mathcal{N}}$, $\max_{i\in\mathcal{N}} |v_i| < 1$ and

$$\sum_{\{i,j\}\in\mathcal{E}} (\Lambda_{ij}\Psi_{i,j}(v_i,v_j) + S_{ij}(v))(v_i - v_j)(w_i - w_j) = \int_\Omega \Pi_{\mathcal{T}} w \,\mathrm{d}f, \text{ for any } w \in \mathbb{R}_0^{\mathcal{N}},$$
$$\text{with } S_{ij}(v) = a_0 h_{\mathcal{T}} \frac{|v_i - v_j|^{p-2}}{d_{ij}^p} \sum_{K\in\mathcal{T}_{ij}} |K| \sum_{k\in\mathcal{N}_K} \psi'(v_k), \text{ for any } \{i,j\} \in \mathcal{E}. \tag{3.8}$$

Remark 3.5. The term $\sum_{(i,j)\in\mathcal{E}} S_{ij}(v)(v_i - v_j)(w_i - w_j)$ can be seen as a discrete p-Laplace stabilisation term since, interpreting each $\frac{v_i - v_j}{d_{ij}}$ as a gradient, it behaves as $a_0 h_{\mathcal{T}} \int_\Omega \psi'(v)|\nabla v|^{p-2}\nabla v \cdot \nabla w \mathrm{d}x$. We notice that this stabilisation term vanishes as $h_{\mathcal{T}} \to 0$, since the definition of the term $S_{ij}(v)$ implies that the term $B_{20}^{(m)}$ in the proof of Theorem 3.7 behaves as $h_{\mathcal{T}}^{1/p}$, thus decreasing the order of convergence if $a_0 > 0$. Note that, although we let $a_0 = 0$ in the numerical results given in this paper, we observe suboptimal convergence orders, resulting from the upstream weighting scheme used in the definition of $\Psi_{i,j}(v_i, v_j)$.

3.3. Main results of Section 3

The first main result of this section is similar to that of Section 2.

Theorem 3.6. *There exists (at least) one solution $v \in \mathbb{R}_0^{\mathcal{N}}$ to Scheme (3.8) (in the sense of Def. 3.4).*

Therefore we can consider a sequence $\left(\mathcal{T}^{(m)}\right)_{m\geq 1}$ of meshes of Ω in the sense specified in Section 2.2 such that (2.7) and (2.8) are satisfied. For this sequence, owing to Theorem 3.6, we can consider for all $m \geq 1$ a solution $v^{(m)}$ to Scheme (3.8) with respect to $\mathcal{T}^{(m)}$. The second main result of this section, also similar to that of Section 2, is that the functions reconstructed from this sequence converge, up to a subsequence, to a weak solution of the continuous problem in the sense of Definition 1.1, as stated by the following theorem.

Theorem 3.7. *Let $\left(\mathcal{T}^{(m)}\right)_{m\geq 1}$ be a sequence of meshes of the computational domain Ω in the sense specified in Section 2.2 such that (2.7) and (2.8) are satisfied. Let $a_0 \geq 0$ and $p \in (2, +\infty)$ be given. For any $m \geq 1$, let $v^{(m)}$ be an arbitrary numerical solution to Scheme (3.8) in the sense of Definition 3.4 in the case where $\mathcal{T} = \mathcal{T}^{(m)}$ and let $u^{(m)} = \psi\left(v^{(m)}\right)$ (the existence of $v^{(m)}$ is given by Thm. 3.6). Then, there exists $\overline{u} \in \mathcal{S}_d(\Omega)$, weak solution of problems (1.1) and (1.2) in the sense of Definition 1.1, such that, up to the extraction of a subsequence, $\left(\Pi_{\mathcal{T}^{(m)}} u^{(m)}\right)_{m\geq 1}$ weakly converges to $\overline{u}$ in $W_0^{1,r}(\Omega)$ for all $r \in \left(1, \frac{d}{d-1}\right)$, and the sequences $\left(\Pi_{\mathcal{T}^{(m)}} u^{(m)}\right)_{m\geq 1}$ and $\left(\Pi_{\mathcal{M}^{(m)}} u^{(m)}\right)_{m\geq 1}$ strongly converge to $\overline{u}$ in $L^q(\Omega)$ for all $q \geq 1$ is $d = 2$ and for all $q \in [1, d/(d-2))$ if $d > 2$.*

Moreover, for all $k > 0$, $T_k\overline{u} \in H_0^1(\Omega)$ holds.

In the case $d = 2$, the whole sequence converges to the unique solution of the problem.

Remark 3.8. Note also that, as in Section 2, $\overline{u} = \psi(v)$ where v is the limit of the subsequence $\left(\Pi_{\mathcal{T}^{(m)}} v^{(m)}\right)_{m\geq 1}$ in $L^q(\Omega)$, for all $q \geq 1$ if $d = 2$ and for all $q \in [1, 2d/(d-2))$ if $d > 2$.

Finally, we have the following important property in the case $f \in L^1(\Omega)$, which is not proved for the scheme studied in Section 2.

Theorem 3.9. *Under the hypotheses and conclusions of Theorem 3.7, moreover assuming that $f \in L^1(\Omega)$, $a_0 \neq 0$ and $p \in (3, +\infty)$, then $\overline{u}$ is the unique weak entropy solution of problems (1.1) and (1.2) in the sense of Definition 3.1, and the whole sequence converges in the sense specified in Theorem 3.7. Moreover, $\Pi_{\mathcal{T}^{(m)}} u^{(m)}$ converges to $\overline{u}$ in $W_0^{1,r}(\Omega)$ for all $r \in \left(1, \frac{d}{d-1}\right)$.*

The remaining part of this section is dedicated to the proof of Theorems 3.6, 3.7 and 3.9.

3.4. Estimates

As in Section 2, we use the function β defined by (2.10) (also represented in the left part of Fig. 3). We define for any $q \in [0, 1)$ the functions $\psi_q : (-1, 1) \to \mathbb{R}$ and $\widetilde{\psi}_q : (-1, 1) \longrightarrow \mathbb{R}$, for any $s \in (-1, 1)$ (represented in the left part of Fig. 3), by

$$\psi_q(s) = \begin{cases} \psi(s) & \text{if } |s| \leq q, \\ \psi(q) + \psi'(q)(s-q) & \text{if } s > q, \\ -\psi(q) + \psi'(q)(s+q) & \text{if } s < -q, \end{cases} \quad \text{and } \widetilde{\psi}_q(s) = \int_0^s \sqrt{\psi'(t)}\sqrt{\psi'_q(t)}\,\mathrm{d}t. \tag{3.9}$$

Note that we have for any $s \in (-1, 1)$,

$$\widetilde{\psi}_q(s) = \begin{cases} \psi(s) & \text{if } |s| \leq q, \\ \psi(q) + \sqrt{\psi'(q)}(\beta(s) - \beta(q)) & \text{if } s > q, \\ -\psi(q) + \sqrt{\psi'(q)}(\beta(s) + \beta(q)) & \text{if } s < -q. \end{cases} \tag{3.10}$$

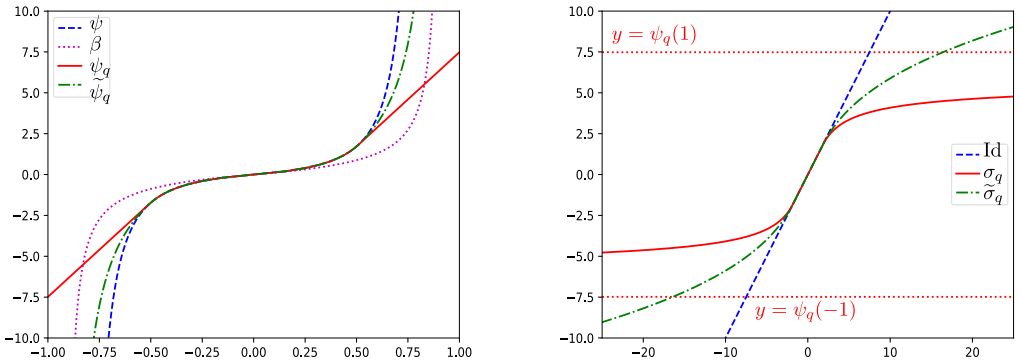

FIGURE 3. *Left*: functions ψ, β, ψ_q and $\widetilde{\psi}_q$ with $q = \frac{1}{2}$. *Right*: functions σ_q and $\widetilde{\sigma}_q$ with $q = \frac{1}{2}$. The lines $y = \psi_q(\pm 1)$ are the asymptotes of the function σ_q at $\pm\infty$.

We finally define the functions $\sigma_q : \mathbb{R} \to \mathbb{R}$ and $\widetilde{\sigma}_q : \mathbb{R} \longrightarrow \mathbb{R}$, for any $s \in \mathbb{R}$ (represented in the right part of Fig. 3), by

$$\sigma_q(s) = \psi_q\big(\psi^{-1}(s)\big) \text{ and } \widetilde{\sigma}_q(s) = \widetilde{\psi}_q\big(\psi^{-1}(s)\big). \tag{3.11}$$

Note that, for all $|s| \le \psi(q)$, $\sigma_q(s) = \widetilde{\sigma}_q(s) = s$.

We have the following estimate.

Lemma 3.10. *Let $v \in \mathbb{R}_0^{\mathcal{N}}$ be such that*

$$\max_{i\in\mathcal{N}} |v_i| < 1 \quad \text{and} \quad \sum_{\{i,j\}\in\mathcal{E}} \Lambda_{ij}\Psi_{i,j}(v_i, v_j)(v_i - v_j)(v_i - v_j) \le \int_\Omega \Pi_{\mathcal{T}} v \,\mathrm{d}f. \tag{3.12}$$

Then we can write

$$\|\nabla \Pi_{\mathcal{T}} \beta(v)\|_2 \le \left(\frac{\|f\|_{M(\Omega)}}{\underline{\lambda}}\right)^{1/2}. \tag{3.13}$$

As a consequence, for any $r \in [1, d/(d-2))$ if $d > 2$ and for any $r \in [1, +\infty)$ if $d = 2$, and for any $\gamma \in [0, +\infty)$ there exists $C_4^{(r,\gamma)}$, which also depends on $\underline{\lambda}$, $\|f\|_{M(\Omega)}$, d and Ω such that

$$\left\| \Pi_{\mathcal{M}} \frac{\psi'(v)}{(1-|v|)^\gamma} \right\|_r \le C_4^{(r,\gamma)}. \tag{3.14}$$

Proof. On one hand, using that $\max_{i\in\mathcal{N}} |v_i| < 1$, we have

$$\int_\Omega \Pi_{\mathcal{T}} v \,\mathrm{d}f \le \|f\|_{M(\Omega)}.$$

On the other hand, for any $\{i, j\} \in \mathcal{E}$, since $\beta(v_i) - \beta(v_j) = \sqrt{\psi'(v_{ij})}(v_i - v_j)$ with $v_{ij} \in I(v_i, v_j)$, we can write

$$\Lambda_{ij}(\beta(v_i) - \beta(v_j))^2 + \Gamma_{ij} = \Lambda_{ij}\Psi_{ij}(v_i, v_j)(v_i - v_j)^2,$$

with

$$\Gamma_{ij} = \Lambda_{ij}(\Psi_{ij}(v_i, v_j) - \psi'(v_{ij}))(v_i - v_j)^2 \ge 0.$$

(Recall that if $\Lambda_{ij} \ge 0$, then $\Psi_{ij}(v_i, v_j) \ge \psi'(v_{ij})$, and if $\Lambda_{ij} \le 0$, then $\Psi_{ij}(v_i, v_j) \le \psi'(v_{ij})$.) Hence we get

$$\underline{\lambda}\|\nabla \Pi_{\mathcal{T}} \beta(v)\|_2^2 \le \int_\Omega \Lambda \nabla \Pi_{\mathcal{T}} \beta(v) \cdot \nabla \Pi_{\mathcal{T}} \beta(v) \,\mathrm{d}x = \sum_{\{i,j\}\in\mathcal{E}} \Lambda_{ij}(\beta(v_i) - \beta(v_j))^2 \le \|f\|_{M(\Omega)},$$

which concludes the proof of (3.13). Let us now turn to the proof of (3.14). For any $\varepsilon \in (0,1)$ and $s \in (-1,1)$, we have

$$\frac{\psi'(s)}{(1-|s|)^\gamma} = \frac{1}{(1-|s|)^{\gamma+2}} \exp \frac{|s|}{1-|s|} \leq \left(\frac{2+\gamma}{\varepsilon}\right)^{2+\gamma} e^{\varepsilon-(2+\gamma)} \exp \frac{(1+\varepsilon)|s|}{1-|s|}.$$

We deduce from (2.13) in the proof of Lemma 2.6 that

$$\exp \frac{(1-\varepsilon)|s|}{2(1-|s|)} \leq \frac{1-\varepsilon}{2} C_\varepsilon |\beta(s)| + 1.$$

Gathering the above results and denoting by $\mu := \frac{2(1+\varepsilon)}{1-\varepsilon}$, we obtain that

$$\forall s \in (-1,1), \ \left(\frac{\psi'(s)}{(1-|s|)^\gamma}\right)^r \leq a_1 |\beta(s)|^{r\mu} + a_2, \tag{3.15}$$

where $a_1 > 0$ and $a_2 > 0$ are only depending on ε, γ and r thanks to Hölder's inequality.

– In the case $d = 2$, let us define $\varepsilon = \frac{1}{2}$. Then a Sobolev inequality (see Lem. A.1) and the equivalence of norms (A.1) provide

$$\|\Pi_{\mathcal{M}} \beta(v)\|_{r\mu} \leq C_{\rm sob}^{(2,r\mu)} C_7^{(r\mu)} \|\nabla \Pi_{\mathcal{T}} \beta(v)\|_2.$$

– In the case $d > 2$, let us select $\varepsilon \in (0,1)$ such that

$$2\frac{1+\varepsilon}{1-\varepsilon} r \leq \frac{2d}{d-2}, \text{ which means that } \frac{1+\varepsilon}{1-\varepsilon} \leq \frac{d}{r(d-2)}.$$

This is possible, since $a_r := \frac{d}{r(d-2)} \in (1, \frac{d}{d-2}]$. It suffices to choose $\varepsilon = \frac{a_r - 1}{a_r + 1} \in (0,1)$ for obtaining $r\mu \leq \frac{2d}{d-2}$.Then a Sobolev inequality and the equivalence of norms (A.1) lead to

$$\|\Pi_{\mathcal{M}} \beta(v)\|_{r\mu} \leq C_{\rm sob}^{(2,r\mu)} C_7^{(r\mu)} \|\nabla \Pi_{\mathcal{T}} \beta(v)\|_2.$$

The above relation and (3.15) yield (3.14). □

Remark 3.11. We get from the previous proof that $\sum_{\{i,j\}\in\mathcal{E}} \Gamma_{ij}$ remains bounded, where Γ_{ij} behaves as a weak 3-Laplace stabilisation. Indeed, such a stabilisation would involve a term $\int_\Omega |\nabla v| \nabla v \cdot \nabla w \mathrm{d}x$ whose discrete version, when $w = v$, behaves as $\sum_{(i,j)\in\mathcal{E}} (m_i + m_j) \frac{(v_i - v_j)^3}{d_{ij}^3}$. In the term Γ_{ij}, since $\Psi_{ij}(v_i, v_j) - \psi'(v_{ij})$ behaves as $v_i - v_j$ and Λ_{ij} behaves as $\frac{m_i + m_j}{d_{ij}^2}$, the sum of terms Γ_{ij} happens to behave as $h \int_\Omega |\nabla v| \nabla v \cdot \nabla w \mathrm{d}x$. Nevertheless, we need a greater stabilisation in the convergence proof to the entropy solution.

As a consequence of the previous result, we can obtain a bound away from one on a function $|\Pi_{\mathcal{T}} v|$ such that the estimate (3.12) holds.

Lemma 3.12. *Let $v \in \mathbb{R}_0^{\mathcal{N}}$ be such that (3.12) holds. Then there exists $C_5 \in (0,1)$ depending only on $\underline{\lambda}, \|f\|_{M(\Omega)}, d, |\Omega|$ and on $\mathcal{T}$ such that*

$$\max_{i\in\mathcal{N}} |v_i| \leq C_5.$$

Proof. Owing to Lemma A.1 with $r = 1$ and to (A.1), we can write

$$\|\Pi_{\mathcal{M}} \beta(v)\|_1 \leq C_7^{(1)} C_{\rm sob}^{(2,1)} \|\nabla \Pi_{\mathcal{T}} \beta(v)\|_2.$$

Consequently using Lemma 3.10 we obtain

$$\sum_{i\in\mathcal{N}} m_i |\beta(v_i)| \leq C_7^{(1)} C_{\rm sob}^{(2,1)} \left(\frac{\|f\|_{M(\Omega)}}{\underline{\lambda}}\right)^{1/2},$$

which gives in particular for any $i \in \mathcal{N}$,

$$|\beta(v_i)| \le \frac{C_7^{(1)} C_{\rm sob}^{(2,1)}}{\min_{i\in\mathcal{N}_{\rm int}} m_i} \left(\frac{\|f\|_{M(\Omega)}}{\underline{\lambda}} \right)^{1/2}.$$

Using the fact that the function β is a continuous strictly increasing one-to-one function from $(-1,1)$ to $\mathbb{R}$, we then obtain for any $i \in \mathcal{N}_{\rm int}$,

$$|v_i| \le C_5 := \beta^{-1}\left(\frac{C_7^{(1)} C_{\rm sob}^{(2,1)}}{\min_{i\in\mathcal{N}_{\rm int}} m_i} \left(\frac{\|f\|_{M(\Omega)}}{\underline{\lambda}} \right)^{1/2} \right) < 1.$$

□

Remark 3.13. Note that the proof of Lemma 3.12 is shorter and simpler than the proof of Lemma 2.8, owing to the finite volume point of view allowed by the CVFE scheme.

Lemma 3.14. *We define, for any $1 \le r < \infty$, the following norm on $\mathbb{R}_0^{\mathcal{N}}$:*

$$\|v\|_{1,r,\mathcal{M}}^r = \sum_{\{i,j\}\in\mathcal{E}} m_{ij} \left| \frac{v_i - v_j}{d_{ij}} \right|^r. \tag{3.16}$$

Let $r \in [1, d/(d-1))$ and let $v \in \mathbb{R}_0^{\mathcal{N}}$ be such that (3.12) holds. Then, denoting by $u = \psi(v)$, there exists C_6 only depending on r, d, $\underline{\lambda}$, $\|f\|_{M(\Omega)}$ and increasingly on $\theta_{\mathcal{T}}$ such that

$$\|u\|_{1,r,\mathcal{M}} \le C_6. \tag{3.17}$$

Proof. We have for any $\{i,j\} \in \mathcal{E}$, using $\max_{s\in I(v_i,v_j)} \sqrt{\psi'(s)} = \max\Big(\sqrt{\psi'(v_i)}, \sqrt{\psi'(v_j)}\Big)$,

$$|\psi(v_i) - \psi(v_j)| = \left| \int_{v_i}^{v_j} \sqrt{\psi'(s)}\sqrt{\psi'(s)}\,\mathrm{d}s \right| \le |\beta(v_i) - \beta(v_j)| \Big(\sqrt{\psi'(v_i)} + \sqrt{\psi'(v_j)} \Big),$$

which gives

$$|u_i - u_j|^r \le 2^{r-1} |\beta(v_i) - \beta(v_j)|^r \Big(\psi'(v_i)^{\frac{r}{2}} + \psi'(v_j)^{\frac{r}{2}} \Big).$$

Summing over the edges, we then obtain

$$\sum_{\{i,j\}\in\mathcal{E}} m_{ij} \left| \frac{u_i - u_j}{d_{ij}} \right|^r \le 2^{r-1} \sum_{\{i,j\}\in\mathcal{E}} m_{ij} \left| \frac{\beta(v_i) - \beta(v_j)}{d_{ij}} \right|^r \Big(\psi'(v_i)^{\frac{r}{2}} + \psi'(v_j)^{\frac{r}{2}} \Big).$$

Using Hlder's inequality with conjugate exponents $\frac{2}{r}$ and $\frac{2}{2-r}$ gives

$$\|u\|_{1,r,\mathcal{M}}^r \le 2^{\frac{3r-2}{2}} \|\beta(v)\|_{1,2,\mathcal{M}}^r \left(\sum_{\{i,j\}\in\mathcal{E}} m_{ij} \Big(\psi'(v_j)^{\frac{r}{2-r}} + \psi'(v_j)^{\frac{r}{2-r}} \Big) \right)^{\frac{2-r}{2}}.$$

Recall that owing to Lemma 3.10 and to (A.2) in Lemma A.2, we have

$$\|\beta(v)\|_{1,2,\mathcal{M}}^2 \le (C_8^{(\theta_{\mathcal{T}},2)})^2 \frac{\|f\|_{M(\Omega)}}{\underline{\lambda}}.$$

Note that, referring to the geometrical definitions of m_{ij} and m_i, we have

$$\sum_{\{i,j\}\in\mathcal{E}} m_{ij}\Big(\psi'(v_j)^{\frac{r}{2-r}}+\psi'(v_j)^{\frac{r}{2-r}}\Big)=\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}\setminus\{i\}}\sum_{K\in\mathcal{T}}\frac{2|K|}{d(d+1)}1_{K\in\mathcal{T}_i}1_{K\in\mathcal{T}_j}\psi'(v_i)^{\frac{r}{2-r}}=2\sum_{i\in\mathcal{N}}\psi'(v_i)^{\frac{r}{2-r}}m_i.$$

This provides

$$\|u\|_{1,r,\mathcal{M}}^r\le 2^{\frac{3r-2}{2}}2^{\frac{2-r}{2}}\|\beta(v)\|_{1,2,\mathcal{M}}^r\|\Pi_{\mathcal{M}}\psi'(v)\|_{\frac{r}{2-r}}^{\frac{r}{2}}=\Big(2\|\beta(v)\|_{1,2,\mathcal{M}}\|\Pi_{\mathcal{M}}\psi'(v)\|_{\frac{r}{2-r}}^{1/2}\Big)^r.$$

Since, for $d=2$, we have $r\in[1,2)$ and therefore $r/(2-r)\in[1,+\infty)$, and for $d>2$, $r<d/(d-1)$ implies $r/(2-r)<d/(d-2)$, we can apply (3.14) in Lemma 3.10, which provides

$$\|\Pi_{\mathcal{M}}\psi'(v)\|_{\frac{r}{2-r}}\le C_4^{(\frac{r}{2-r},0)}.$$

Gathering the preceding inequalities provides the conclusion of the lemma. □

Lemma 3.15. *Let $v\in\mathbb{R}_0^{\mathcal{N}}$ be a solution of the numerical scheme (3.8), and let $u=\psi(v)$. Then, for any $q\in(0,1)$, we have*

$$\|\nabla\Pi_{\mathcal{T}}\widetilde{\sigma}_q(u)\|_2\le\frac{\|f\|_{M(\Omega)}}{\underline{\lambda}}(\psi(q)+\psi'(q)(1-q)).\tag{3.18}$$

Proof. We take $\psi_q(v)$ in the numerical scheme (3.8) and we obtain

$$\sum_{\{i,j\}\in\mathcal{E}}\Lambda_{ij}\Psi_{ij}(v_i,v_j)(v_i-v_j)(\psi_q(v_i)-\psi_q(v_j))\le\int_\Omega\Pi_{\mathcal{T}}\psi_q(v)\,\mathrm{d}f.$$

Using the definition of ψ_q we have

$$\int_\Omega\Pi_{\mathcal{T}}\psi_q(v)\,\mathrm{d}f\le\|f\|_{\mathcal{M}(\Omega)}(\psi(q)+\psi'(q)(1-q)).$$

Using Lemma A.4, we obtain for $v_i\neq v_j$,

$$\min_{s\in I(v_i,v_j)}\psi'(s)\le\frac{\big(\widetilde{\psi}_q(v_i)-\widetilde{\psi}_q(v_j)\big)^2}{(v_i-v_j)(\psi_q(v_i)-\psi_q(v_j))}\le\max_{s\in I(v_i,v_j)}\psi'(s).$$

This yields

$$\int_\Omega\Lambda\nabla\Pi_{\mathcal{T}}\widetilde{\psi}_q(v)\cdot\nabla\Pi_{\mathcal{T}}\widetilde{\psi}_q(v)\,\mathrm{d}x=\sum_{\{i,j\}\in\mathcal{E}}\Lambda_{ij}\Big(\widetilde{\psi}_q(v_i)-\widetilde{\psi}_q(v_j)\Big)^2\le\|f\|_{\mathcal{M}(\Omega)}(\psi(q)+\psi'(q)(1-q)),$$

By definition (3.11) of $\widetilde{\sigma}_q$, we have the equality $\widetilde{\sigma}_q(u)=\widetilde{\psi}_q(v)$, which provides the conclusion. □

Remark 3.16. One cannot directly get an estimate on $T_k(u)$ by taking $T_k(\psi(v))$ as test function in (3.8). Indeed, letting $q=\psi^{-1}(k)$, this would request that, for any $(a,b)\in(-1,1)^2$ with $T_q(a)\neq T_q(b)$, it holds

$$\min_{s\in I(a,b)}\psi'(s)\le\frac{(\psi(T_q(b))-\psi(T_q(a)))^2}{(b-a)(\psi(T_q(b))-\psi(T_q(a)))}\le\max_{s\in I(a,b)}\psi'(s).$$

But, although the right above inequality holds, this is not the case for the left one: considering $0<a<q<b$ and letting a tend to q, then $\frac{(\psi(q)-\psi(a))^2}{(b-a)(\psi(q)-\psi(a))}$ tends to 0 whereas $\min_{s\in I(a,b)}\psi'(s)\ge 1$. Recall that the left side, only used in the case $\Lambda_{ij}<0$, is not used in finite volume-type approaches.

Lemma 3.17 (Weak p-Laplace inequality)**.** *Let $v \in \mathbb{R}_0^{\mathcal{N}}$ be such that Scheme (3.8) holds. Then it holds*

$$a_0 h_{\mathcal{T}} \sum_{\{i,j\}\in\mathcal{E}} \frac{(v_i - v_j)^p}{d_{ij}^p} \sum_{K\in\mathcal{T}_{ij}} |K| \sum_{k\in\mathcal{N}_K} \psi'(v_k) \le \|f\|_{M(\Omega)}. \tag{3.19}$$

Proof. We let v as test function in the numerical scheme (3.8) and we obtain

$$\sum_{\{i,j\}\in\mathcal{E}} \Big(\Lambda_{ij}\Psi_{i,j}(v_i,v_j)(v_i-v_j)^2 + S_{ij}(v)\Big)(v_i-v_j)^2 = \int_\Omega \Pi_{\mathcal{T}} v \,\mathrm{d}f.$$

Since the proof of Lemma 3.10 provides $\sum_{\{i,j\}\in\mathcal{E}} \Lambda_{ij}\Psi_{i,j}(v_i,v_j)(v_i-v_j)^2 \ge 0$, we get

$$\sum_{\{i,j\}\in\mathcal{E}} S_{ij}(v)(v_i-v_j)^2 \le \int_\Omega \Pi_{\mathcal{T}} v \,\mathrm{d}f \le \|f\|_{M(\Omega)},$$

which yields, introducing the expression of $S_{ij}(v)$,

$$\sum_{\{i,j\}\in\mathcal{E}} a_0 h_{\mathcal{T}} \frac{(v_i - v_j)^p}{d_{ij}^p} \sum_{K\in\mathcal{T}_{ij}} |K| \sum_{k\in\mathcal{N}_K} \psi'(v_k) \le \|f\|_{M(\Omega)}.$$

This is (3.19). □

3.5. Existence of a solution to CVFE Scheme (3.8)

The purpose of this section is to prove Theorem 3.6, which states the existence of a solution to the numerical scheme in the sense of Definition 3.4, by applying the topological degree method [12]. This proof follows the same lines as that of Theorem 2.3.

Proof of Theorem 3.6. We define the canonical basis $(\delta^k)_{k\in\mathcal{N}_{\rm int}}$ of $\mathbb{R}_0^{\mathcal{N}}$, by setting $\delta_i^k = 1$ if $i = k$ and 0 otherwise, for any $i \in \mathcal{N}$. Let us define the continuous function

$$\mathcal{F}: \quad \left| \begin{array}{l} \mathbb{R}_0^{\mathcal{N}} \times [0,1] \longrightarrow \mathbb{R}_0^{\mathcal{N}} \\ (u,\mu) \longmapsto \mathcal{F}(u,\mu) = (\mathcal{F}_k(u,\mu))_{k\in\mathcal{N}}, \end{array} \right. \tag{3.20}$$

where for any $u = (u_i)_{i\in\mathcal{N}}$, $\mu \in [0,1]$ and for any $k \in \mathcal{N}_{\rm ext}$, $\mathcal{F}_k(u,\mu) := 0$ and for any $k \in \mathcal{N}_{\rm int}$, the function $\mathcal{F}_k(u,\mu)$ is defined by

$$\mathcal{F}_k(u,\mu) := \mu\left(\sum_{\{i,j\}\in\mathcal{E}} (\Lambda_{ij}\Psi_{i,j}(v_i,v_j) + S_{ij}(v))(v_i-v_j)(\delta_i^k - \delta_j^k) - \int_\Omega \Pi_{\mathcal{T}}\delta^k \,\mathrm{d}f\right) + (1-\mu)u_k,$$

where for all $i \in \mathcal{N}$, we denote by $v_i = \psi^{-1}(u_i)$ (notice that $\Pi_{\mathcal{T}}\delta^k = \varphi_k$). This mapping is well defined and continuous, since, for any $u = (u_i)_{i\in\mathcal{N}} \in \mathbb{R}_0^{\mathcal{N}}$, we have $\max_{i\in\mathcal{N}} |\psi^{-1}(u_i)| < 1$. We also notice that the equation $\mathcal{F}(u,1) = 0$ is equivalent to state that $v = \psi^{-1}(u) \in \mathbb{R}_0^{\mathcal{N}}$ is a solution to Scheme (3.8). Let $\mu \in (0,1]$ and let $u = (u_i)_{i\in\mathcal{N}} \in \mathbb{R}_0^{\mathcal{N}}$ be such that $\mathcal{F}(u,\mu) = 0$. Multiplying $\mathcal{F}_k(u,\mu)$ by $\psi^{-1}(u_k) = v_k$ and summing on $k \in \mathcal{N}_{\rm int}$, we obtain

$$\mu\left(\sum_{\{i,j\}\in\mathcal{E}} (\Lambda_{ij}\Psi_{i,j}(v_i,v_j) + S_{ij}(v))(v_i-v_j)^2 - \int_\Omega \Pi_{\mathcal{T}} v \,\mathrm{d}f\right) + (1-\mu)\sum_{k\in\mathcal{N}_{\rm int}} u_k\psi^{-1}(u_k) = 0.$$

This implies, since $\mu \in (0,1]$ and $S_{ij}(v) \geq 0$, that (3.12) holds for v. Hence, from Lemma 3.12, we obtain

$$\max_{i\in\mathcal{N}} |v_i| \leq C_5 < 1,$$

leading to

$$|u_i| \leq \psi(C_5), \ \forall i \in \mathcal{N}.$$

Define the relatively compact open set

$$\mathcal{U} = \{u = (u_i)_{i\in\mathcal{N}} \in \mathbb{R}_0^{\mathcal{N}} \text{ such that } |u_i| < \psi(C_5) + 1 \text{ for all } i \in \mathcal{N}\}.$$

For $\mu = 0$, the linear equation $\mathcal{F}(u,0) = 0$ has the unique solution $u = 0$. The topological degree corresponding to $\mathcal{F}(u,0)$ and $\mathcal{U}$ is therefore equal to 1 since $u = 0$ belongs to $\mathcal{U}$. Hence for any $\mu \in [0,1]$ any solution to $\mathcal{F}(u,\mu) = 0$ necessarily belongs to $\mathcal{U}$. Therefore, owing to the invariance of the topological degree by homotopy, there exists at least one $u \in \mathbb{R}_0^{\mathcal{N}}$ (not necessarily unique) such that $\mathcal{F}(u,1) = 0$, which means that $v = \psi^{-1}(u)$ is a solution to Scheme (3.8) in the sense of Definition 3.4. □

3.6. Convergence to a weak solution in the general case $f \in M(\Omega)$

The goal of this section is the proof of Theorem 3.7. In this section, it is possible to let $a_0 = 0$, which means that the convergence results to a weak solution also hold without the stabilisation term. We could as well consider $p \in (1,2]$ in the stabilisation term under a suitable definition of $S_{ij}(v)$ in the case $v_i = v_j$. The first step is the following compactness lemma.

Lemma 3.18. *Let Ω be an open bounded subset of $\mathbb{R}^d$ with $d \geq 2$. Let $\left(\mathcal{T}^{(m)}\right)_{m\geq 0}$ be a sequence of simplicial meshes in the sense of Section 3.2, such that (2.7) and (2.8) hold. For any $m \geq 0$, let $v^{(m)}$ be a solution to Scheme (3.8) and let $u^{(m)} = \psi\left(v^{(m)}\right)$. Then there exist $\overline{u} \in \mathcal{S}_d(\Omega)$ and a subsequence of $\left(\mathcal{T}^{(m)}, v^{(m)}\right)_{m\geq 0}$, again denoted $\left(\mathcal{T}^{(m)}, v^{(m)}\right)_{m\geq 0}$, such that:*

(1) *for all $r \in [1, d/(d-1))$, $\nabla \Pi_{\mathcal{T}^{(m)}} u^{(m)}$ weakly converges to $\nabla \overline{u}$ in $L^r(\Omega)^d$,*
(2) *for all $s \in [1, +\infty)$ if $d = 2$ and for all $s \in [1, d/(d-2))$ if $d > 2$, $\Pi_{\mathcal{T}^{(m)}} u^{(m)}$ and $\Pi_{\mathcal{M}^{(m)}} u^{(m)}$ converge to $\overline{u}$ in $L^s(\Omega)$ and almost everywhere in Ω,*
(3) *for all $q \in (0,1)$, $\nabla \Pi_{\mathcal{T}^{(m)}} \widetilde{\sigma}_q\left(u^{(m)}\right)$ weakly converges to $\nabla \widetilde{\sigma}_q(\overline{u})$ in $L^2(\Omega)$ and $\widetilde{\sigma}_q(\overline{u}) \in H_0^1(\Omega)$ which implies that $T_k(\overline{u}) \in H_0^1(\Omega)$ for all $k > 0$.*

Proof. Let us denote the initial sequence by $\left(\mathcal{T}^{(m)}, v^{(m)}\right)_{m\in\mathbb{N}}$. Thanks to (3.17) in Lemma 3.14 and to (A.2) in Lemma A.2, for a given $r_0 \in \left(1, \frac{d}{d-1}\right)$, we can select $\overline{u} \in W_0^{1,r_0}(\Omega)$ and an infinite subset $S_1 \subset \mathbb{N}$ such that $\left(\Pi_{\mathcal{T}^{(m)}} u^{(m)}\right)_{m\in S_1}$ weakly converges to $\overline{u}$ in $W_0^{1,r_0}(\Omega)$ and converges almost everywhere in Ω.

Let $r \in (1, d/(d-1))$ be given. Owing again to the same arguments, we deduce that there exists $\overline{u}' \in W_0^{1,r}(\Omega)$ and an infinite subset $S_2 \subset S_1$ such that $\left(\Pi_{\mathcal{T}^{(m)}} u^{(m)}\right)_{m\in S_2}$ weakly converges to $\overline{u}'$ in $W_0^{1,r}(\Omega)$. Then $\overline{u} = \overline{u}'$ in $W_0^{1,\min(r,r_0)}(\Omega)$, which implies, by uniqueness of the limit, that in fact we can take $S_2 = S_1$ and that we have $\overline{u} \in W_0^{1,\max(r,r_0)}(\Omega)$. Since this holds for all $r \in (1, d/(d-1))$, we get that $\overline{u} \in \mathcal{S}_d(\Omega)$ and that the sequence $\left(\mathcal{T}^{(m)}, v^{(m)}\right)_{m\in S_1}$ satisfies the weak convergence property for any $r \in (1, d/(d-1))$. This concludes the first item.

The second item is a direct consequence of Sobolev inequalities for the convergence properties of $\left(\Pi_{\mathcal{T}^{(m)}} u^{(m)}\right)_{m\in S_1}$ in $L^s(\Omega)$ and almost everywhere in Ω. Then the application of (A.1) in Lemma A.2 and of Lemma A.3 provides the same conclusion for $\left(\Pi_{\mathcal{M}^{(m)}} u^{(m)}\right)_{m\in S_1}$.

We then get, remarking that $\widetilde{\sigma}_q\left(\Pi_{\mathcal{M}^{(m)}} u^{(m)}\right) = \Pi_{\mathcal{M}^{(m)}} \widetilde{\sigma}_q\left(u^{(m)}\right)$ (important property of the piecewise functions) that $\left(\Pi_{\mathcal{M}^{(m)}} \widetilde{\sigma}_q\left(u^{(m)}\right)\right)_{m\in S_1}$ converges to $\widetilde{\sigma}_q(\overline{u})$ in $L^2(\Omega)$, and applying again Lemma A.3 in addition

to Lemma 3.15, we get that $\left(\Pi_{\mathcal{T}^{(m)}}\widetilde{\sigma}_q\left(u^{(m)}\right)\right)_{m\in S_1}$ converges to $\widetilde{\sigma}_q(\overline{u})$ in $L^2(\Omega)$. This yields, accounting from Lemma 3.15, the weak convergence of $\nabla\Pi_{\mathcal{T}^{(m)}}\widetilde{\sigma}_q\left(u^{(m)}\right)$ to $\nabla\widetilde{\sigma}_q(\overline{u})$ in $L^2(\Omega)$.

We then get that $\widetilde{\sigma}_q(\overline{u}) \in H_0^1(\Omega)$ by considering the convergence of the continuation by 0 outside Ω of $\Pi_{\mathcal{T}^{(m)}}\widetilde{\sigma}_q(u^{(m)})$ in $L^2(\mathbb{R}^d)$ and the weak convergence of its gradient in $L^2(\mathbb{R}^d)$. In order to prove that $T_k(\overline{u}) \in H_0^1(\Omega)$ for any $k>0$,we choose $q\in(0,1)$ such that $\psi(q)>k$. Since we have

$$T_k(\overline{u}) = T_k(\widetilde{\sigma}_q(\overline{u})),$$

we get that, applying Stampacchia's results [25],

$$\int_\Omega |\nabla T_k(\overline{u})|^2\,\mathrm{d}x = \int_\Omega |\nabla T_k(\widetilde{\sigma}_q(\overline{u}))|^2\,\mathrm{d}x = \int_\Omega (T_k'(\widetilde{\sigma}_q(\overline{u}))^2 |\nabla\widetilde{\sigma}_q(\overline{u})|^2\,\mathrm{d}x.$$

Since $|T_k'(s)|\le 1$, we deduce that

$$\|\nabla T_k(\overline{u})\|_2 \le \|\nabla\widetilde{\sigma}_q(\overline{u})\|_2,$$

which proves that $T_k\overline{u}\in H_0^1(\Omega)$. □

We now turn to the convergence proof to a weak solution of the continuous problem.

Proof of Theorem 3.7. Applying Lemma 3.18, let us prove that $\overline{u}$ is a weak solution of problems (1.1) and (1.2) in the sense of Definition 1.1.

For all $m\in\mathbb{N}$, we rewrite (2.21) and (2.22) with the notations of CVFE schemes: the interpolation operator

$$\mathcal{P}_{\mathcal{T}^{(m)}}: \left|\begin{array}{l} C_c^\infty(\Omega) \longrightarrow \mathbb{R}_0^{\mathcal{N}^{(m)}} \\ \phi \longmapsto \left(\phi\left(z_i^{(m)}\right)\right)_{i\in\mathcal{N}^{(m)}}, \end{array}\right. \tag{3.21}$$

satisfies the following approximation properties (see [16]): for any $\phi\in C_c^\infty(\Omega)$,

$$\|\phi - \Pi_{\mathcal{T}^{(m)}}\mathcal{P}_{\mathcal{T}^{(m)}}(\phi)\|_\infty + h_{\mathcal{T}^{(m)}}\|\nabla\phi - \nabla\Pi_{\mathcal{T}^{(m)}}\mathcal{P}_{\mathcal{T}^{(m)}}(\phi)\|_\infty \le C_{\mathrm{inter}}^{(\theta_{\mathcal{T}^{(m)}})} h_{\mathcal{T}^{(m)}}^2 \sum_{k=1}^d\sum_{\ell=1}^d \left\|\frac{\partial^2\phi}{\partial x_k\partial x_\ell}\right\|_\infty, \tag{3.22}$$

where $C_{\mathrm{inter}}^{(\theta_{\mathcal{T}^{(m)}})}$ is increasingly depending on $\theta_{\mathcal{T}^{(m)}}$. For a given $\phi\in C_c^\infty(\Omega)$, we let $\mathcal{P}_{\mathcal{T}^{(m)}}(\phi)$ in (3.8). Dropping some indices m and denoting for short ϕ_i instead of $\phi(z_i)$, we get for any $m\ge 1$,

$$B_1^{(m)} + B_2^{(m)} = B_3^{(m)},$$

with

$$\begin{aligned} B_1^{(m)} &:= \sum_{\{i,j\}\in\mathcal{E}} \Lambda_{ij}\Psi_{i,j}(v_i,v_j)(v_i-v_j)(\phi_i-\phi_j), \\ B_2^{(m)} &:= \sum_{\{i,j\}\in\mathcal{E}} S_{ij}(v)(v_i-v_j)(\phi_i-\phi_j), \\ B_3^{(m)} &:= \int_\Omega \Pi_{\mathcal{T}}\mathcal{P}_{\mathcal{T}}(\phi)\,\mathrm{d}f. \end{aligned}$$

From (3.22) and the continuity property of $f\in M(\Omega)$, we get that

$$\lim_{m\to\infty} B_3^{(m)} = \int_\Omega \phi\,\mathrm{d}f.$$

Studying $B_2^{(m)}$, we have the existence of C_ϕ such that $|\phi_i - \phi_j| \le C_\phi d_{ij}$ for all $\{i,j\} \in \mathcal{E}$. Therefore

$$\left|B_2^{(m)}\right| \le B_{20}^{(m)} := C_\phi a_0 h_{\mathcal{T}} \sum_{\{i,j\}\in\mathcal{E}} \frac{(v_i - v_j)^{p-1}}{d_{ij}^{p-1}} \sum_{K\in\mathcal{T}_{ij}} |K| \sum_{k\in\mathcal{N}_K} \psi'(v_k).$$

This provides, thanks to Hölder's inequality with exponents $p/(p-1)$ and p, and using Lemma 3.17,

$$\left|B_{20}^{(m)}\right| \le C_\phi (a_0 h_{\mathcal{T}})^{\frac{1}{p}} \left(\|f\|_{M(\Omega)}\right)^{\frac{p-1}{p}} \left(B_{21}^{(m)}\right)^{\frac{1}{p}},$$

with

$$B_{21}^{(m)} = \sum_{\{i,j\}\in\mathcal{E}} \sum_{K\in\mathcal{T}_{ij}} |K| \sum_{k\in\mathcal{N}_K} \psi'(v_k) = \frac{d(d+1)^2}{2} \|\Pi_{\mathcal{M}} \psi'(v)\|_1 \le \frac{d(d+1)^2}{2} C_4^{(1,0)},$$

owing to (3.14) in Lemma 3.10. Hence we get

$$\lim_{m\to\infty} B_{20}^{(m)} = \lim_{m\to\infty} B_2^{(m)} = 0.$$

Let us now turn to $B_1^{(m)} = B_{11}^{(m)} + B_{12}^{(m)}$, defining

$$B_{11}^{(m)} := \sum_{\{i,j\}\in\mathcal{E}} \Lambda_{ij}(u_i - u_j)(\phi_i - \phi_j) = \int_\Omega \Lambda \nabla \Pi_{\mathcal{T}^{(m)}} u \cdot \nabla \Pi_{\mathcal{T}^{(m)}} \mathcal{P}_{\mathcal{T}^{(m)}}(\phi)\,\mathrm{d}x,$$
$$B_{12}^{(m)} := \sum_{\{i,j\}\in\mathcal{E}} \Lambda_{ij}(\Psi_{i,j}(v_i,v_j) - \psi'(v_{1,ij}))(v_i - v_j)(\phi_i - \phi_j),$$

with

$$v_{1,ij} \in I(v_i, v_j) \text{ such that } u_i - u_j = \psi'(v_{1,ij})(v_i - v_j). \tag{3.23}$$

Owing to (3.22) and (2.8), we get that the sequence $(\Pi_{\mathcal{T}^{(m)}} \mathcal{P}_{\mathcal{T}^{(m)}}(\phi))_{m\ge1}$ converges to ϕ in $L^\infty(\Omega)$ and that the sequence $(\nabla \Pi_{\mathcal{T}^{(m)}} \mathcal{P}_{\mathcal{T}^{(m)}}(\phi))_{m\ge1}$ converges to $\nabla\phi$ in $L^\infty(\Omega)$. Since the sequence $\left(\Pi_{\mathcal{T}^{(m)}} u^{(m)}\right)_{m\ge1}$ weakly converges to $\overline{u}$ in all $W_0^{1,p}(\Omega)$ for all $p \in (1, d/(d-1))$ we then obtain

$$\lim_{m\to\infty} B_{11}^{(m)} = \lim_{m\to\infty} \int_\Omega \Lambda \nabla \Pi_{\mathcal{T}^{(m)}} u \cdot \nabla \Pi_{\mathcal{T}^{(m)}} \mathcal{P}_{\mathcal{T}^{(m)}}(\phi)\,\mathrm{d}x = \int_\Omega \Lambda \nabla \overline{u} \cdot \nabla \phi\,\mathrm{d}x.$$

Turning to the study of $B_{12}^{(m)}$, we have, again using $|\phi_i - \phi_j| \le C_\phi d_{ij}$ and writing $|\Lambda_{ij}| \le \overline{\lambda} \sum_{K\in\mathcal{T}_{ij}} |K| \frac{\theta_{\mathcal{T}}^2}{d_{ij}^2}$,

$$\left|B_{12}^{(m)}\right| \le B_{120}^{(m)} := \overline{\lambda} \sum_{\{i,j\}\in\mathcal{E}} \sum_{K\in\mathcal{T}_{ij}} |K| \frac{\theta_{\mathcal{T}}^2}{d_{ij}^2} |\Psi_{i,j}(v_i,v_j) - \psi'(v_{1,ij})||v_i - v_j| C_\phi d_{ij}.$$

We observe that

$$\Psi_{i,j}(v_i,v_j) = \psi'(v_{2,ij}) \text{ with } v_{2,ij} \in I(v_i,v_j), \tag{3.24}$$

and $|\psi'(v_{2,ij}) - \psi'(v_{1,ij})| \le \max_{s\in I(v_i,v_j)} \psi'(s) - \min_{s\in I(v_i,v_j)} \psi'(s)$. This provides

$$B_{120}^{(m)} \le C_\phi \theta_{\mathcal{T}}^2 \overline{\lambda} \sum_{\{i,j\}\in\mathcal{E}} \sum_{K\in\mathcal{T}_{ij}} |K| \frac{|v_i - v_j|}{d_{ij}} (\psi'(v_i) + \psi'(v_j)) \zeta_{ij}.$$

with

$$\zeta_{ij} := \frac{\max_{s\in I(v_i,v_j)} \psi'(s) - \min_{s\in I(v_i,v_j)} \psi'(s)}{\psi'(v_i) + \psi'(v_j)} \in [0,1).$$

For a value $\varepsilon \in (0, \frac{7}{4})$ such that $0 < \varepsilon < \frac{d}{d-2} - 1$ if $d > 2$, $\varepsilon = \frac{1}{2}$ if $d = 2$, we apply (A.4) in Lemma A.5. We thus get

$$B_{120}^{(m)} \le C_\phi \theta_{\mathcal{T}}^2 \overline{\lambda} \nu_\varepsilon \sum_{\{i,j\}\in\mathcal{E}} \sum_{K\in\mathcal{T}_{ij}} |K| \frac{|\beta(v_i) - \beta(v_j)|}{d_{ij}} (2 + |\beta(v_i)|^{1+\varepsilon} + |\beta(v_j)|^{1+\varepsilon}) \zeta_{ij}.$$

Owing to the Cauchy–Schwarz inequality, we have $\left(B_{120}^{(m)}\right)^2 \le 3\left(C_\phi \theta_{\mathcal{T}}^2 \overline{\lambda} \nu_\varepsilon\right)^2 B_{121}^{(m)} B_{122}^{(m)}$ with

$$B_{121}^{(m)} := \sum_{\{i,j\}\in\mathcal{E}} \sum_{K\in\mathcal{T}_{ij}} |K| \frac{(\beta(v_i) - \beta(v_j))^2}{d_{ij}^2},$$

and

$$B_{122}^{(m)} := \sum_{\{i,j\}\in\mathcal{E}} \sum_{K\in\mathcal{T}_{ij}} |K| \left(4 + |\beta(v_i)|^{2+2\varepsilon} + |\beta(v_j)|^{2+2\varepsilon}\right) \zeta_{ij}^2.$$

We have, applying (A.2) in Lemmas A.2 and 3.10,

$$B_{121}^{(m)} = \frac{d(d+1)}{2} \|\beta(v)\|_{1,2,\mathcal{M}}^2 \le \frac{d(d+1)}{2} \left(C_8^{(\theta_{\mathcal{T}},2)}\right)^2 \|\nabla \Pi_{\mathcal{T}} \beta(v)\|_2^2 \le \frac{d(d+1)}{2} C_8^{(\theta_{\mathcal{T}},2)} \frac{\|f\|_{M(\Omega)}}{\underline{\lambda}}.$$

Turning to $B_{122}^{(m)}$, we have

$$B_{122}^{(m)} = \sum_{i\in\mathcal{N}} \omega_i \left(2 + |\beta(v_i)|^{2+2\varepsilon}\right) \zeta_i,$$

defining for all $i \in \mathcal{N}$,

$$\zeta_i := \frac{1}{\omega_i} \sum_{K\in\mathcal{T}_i} |K| \sum_{j\in\mathcal{N}_K\setminus\{i\}} \zeta_{ij}^2.$$

This yields

$$B_{122}^{(m)} = \int_\Omega \Pi_{\mathcal{M}} \left(2 + |\beta(v)|^{2+2\varepsilon}\right) \Pi_{\mathcal{M}} \zeta \, \mathrm{d}x.$$

We now apply Hölder's inequality with conjugate exponents e_1, e_2, where e_1 is such that $e_1(2+2\varepsilon) = 2d/(d-2)$ if $d > 2$ (hence $e_1 > 1$ owing to the choice of ε), $e_1 = 2$ if $d = 2$, and $e_2 > 1$ is such that $1/e_1 + 1/e_2 = 1$. We obtain

$$B_{122}^{(m)} \le \left(\left\|\Pi_{\mathcal{M}} \left(2 + |\beta(v)|^{2+2\varepsilon}\right)\right\|_{e_1}\right)^{1/e_1} \left(\|\Pi_{\mathcal{M}} \zeta\|_{e_2}\right)^{1/e_2},$$

This choice of e_1 suffices for obtaining that $\left\|\Pi_{\mathcal{M}} \left(2 + |\beta(v)|^{2+2\varepsilon}\right)\right\|_{e_1}$ remains bounded thanks to (A.1) in Lemma A.2, to Lemma 3.10 and to a Sobolev inequality.

Applying Lemma 3.19 below, we get that

$$\lim_{m\to\infty} \left\|\Pi_{\mathcal{M}^{(m)}} \zeta^{(m)}\right\|_{e_2} = 0,$$

which suffices to prove that

$$\lim_{m\to\infty} B_{122}^{(m)} = \lim_{m\to\infty} B_{120}^{(m)} = \lim_{m\to\infty} B_{12}^{(m)} = 0.$$

Consequently, equation (1.5) holds for any $\phi \in C_c^\infty(\Omega)$. By a density argument we obtain that (1.5) holds for any $\phi \in \mathcal{T}_d(\Omega)$, which gives that $\overline{u}$ is a weak solution of problems (1.1) and (1.2) in the sense of Definition 1.1. □

Lemma 3.19. *Under the hypotheses of Theorem 3.7, the function $\Pi_{\mathcal{M}^{(m)}} \zeta^{(m)}$ defined in the proof of this theorem is such that*

$$\forall r \in [1, +\infty), \lim_{m\to\infty} \left\|\Pi_{\mathcal{M}^{(m)}} \zeta^{(m)}\right\|_r = 0.$$

Proof. Using $\psi' \geq 1$, we first remark that, for all $i \in \mathcal{N}$,

$$\zeta_i \leq \frac{1}{\omega_i} \sum_{K \in \mathcal{T}_i} |K| \sum_{j \in \mathcal{N}_K \setminus \{i\}} \left(\max_{s \in I(v_i, v_j)} \psi'(s) - \min_{s \in I(v_i, v_j)} \psi'(s) \right).$$

Since for all $s \in (-1, 0) \cup (0, 1)$, we have $|\psi''(s)| = \frac{3-2|s|}{(1-|s|)^2} \psi'(s)$, we have for any $a < b \in (-1, 1)$

$$\begin{aligned} \max_{s \in I(a,b)} \psi'(s) - \min_{s \in I(a,b)} \psi'(s) &\leq \int_a^b |\psi''(s)| \, \mathrm{d}s \leq 3 \int_a^b \frac{\psi'(s)}{(1-|s|)^2} \, \mathrm{d}s \\ &\leq 3 \int_a^b \sqrt{\psi'(s)} \, \mathrm{d}s \left(\frac{\sqrt{\psi'(a)}}{(1-|a|)^2} + \frac{\sqrt{\psi'(b)}}{(1-|b|)^2} \right) \\ &= 3(\beta(b) - \beta(a)) \left(\frac{\sqrt{\psi'(a)}}{(1-|a|)^2} + \frac{\sqrt{\psi'(b)}}{(1-|b|)^2} \right). \end{aligned}$$

We therefore get, dividing and multiplying by d_{ij}, that

$$\|\Pi_{\mathcal{M}} \zeta\|_1 \leq 3 h_{\mathcal{T}} \sum_{i \in \mathcal{N}} \sum_{K \in \mathcal{T}_i} |K| \sum_{j \in \mathcal{N}_K \setminus \{i\}} \frac{|\beta(v_j) - \beta(v_i)|}{d_{ij}} \left(\frac{\sqrt{\psi'(v_i)}}{(1-|v_i|)^2} + \frac{\sqrt{\psi'(v_j)}}{(1-|v_j|)^2} \right).$$

Applying the Cauchy–Schwarz inequality provides

$$\begin{aligned} \|\Pi_{\mathcal{M}} \zeta\|_1^2 \leq 9 h_{\mathcal{T}}^2 &\left(\sum_{i \in \mathcal{N}} \sum_{K \in \mathcal{T}_i} |K| \sum_{j \in \mathcal{N}_K \setminus \{i\}} \frac{(\beta(v_j) - \beta(v_i))^2}{d_{ij}^2} \right) \\ &\times \left(\sum_{i \in \mathcal{N}} \sum_{K \in \mathcal{T}_i} |K| \sum_{j \in \mathcal{N}_K \setminus \{i\}} \left(\frac{\psi'(v_i)}{(1-|v_i|)^4} + \frac{\psi'(v_j)}{(1-|v_j|)^4} \right) \right). \end{aligned}$$

Using (3.13) in Lemma 3.10 and (A.2) in Lemma A.2, we have

$$\sum_{i \in \mathcal{N}} \sum_{K \in \mathcal{T}_i} |K| \sum_{j \in \mathcal{N}_K \setminus \{i\}} \frac{(\beta(v_j) - \beta(v_i))^2}{d_{ij}^2} = d(d+1) \|\beta(v)\|_{1,2,\mathcal{M}}^2 \leq d(d+1) \left(C_8^{(\theta_{\mathcal{T}}, 2)} \right)^2 \frac{\|f\|_{M(\Omega)}}{\underline{\lambda}},$$

and using (3.14) in Lemma 3.10, we have

$$\sum_{i \in \mathcal{N}} \sum_{K \in \mathcal{T}_i} |K| \sum_{j \in \mathcal{N}_K \setminus \{i\}} \left(\frac{\psi'(v_i)}{(1-|v_i|)^4} + \frac{\psi'(v_j)}{(1-|v_j|)^4} \right) = 2d(d+1) \|\Pi_{\mathcal{M}} \frac{\psi'(v)}{(1-|v|)^4}\|_1 \leq 2d(d+1) C_4^{(1,4)}.$$

The preceding inequalities imply that

$$\lim_{m \to \infty} \left\| \Pi_{\mathcal{M}^{(m)}} \zeta^{(m)} \right\|_1 = 0.$$

Remarking that $\zeta_i \in [0, d(d+1)]$, we have, for $r \geq 1$

$$\|\Pi_{\mathcal{M}} \zeta\|_r^r \leq (d(d+1))^{r-1} \|\Pi_{\mathcal{M}} \zeta\|_1,$$

hence concluding the proof. □

3.7. Convergence to the entropy solution in the case $f \in L^1(\Omega)$

In this section, we assume that the right hand side of (1.1) is defined by $f \in L^1(\Omega)$, and that $a_0 > 0$ and $p > 3$, which are requested in the course of the convergence proof. The next lemma provides a general convergence result due to the presence of the p-Laplace stabilisation term. This result is used several times in the proof of convergence to the entropy solution.

Lemma 3.20. *Let $\left(\mathcal{T}^{(m)}\right)_{m\geq 1}$ be a sequence of simplicial meshes of Ω such that (2.7) and (2.8) hold. Let $(v_{\mathcal{T}^{(m)}})_{m\geq 1}$ be a sequence of solutions to the numerical scheme (3.8) in the sense of Definition 3.4 with $a_0 > 0$ and $p > 3$. Let us define, for any $\alpha \in (0,3]$, $\gamma \in [0,+\infty)$, and for any $m \geq 1$ (dropping some indices m in the discrete quantities involved in the right-hand side),*

$$A^{(m)}(\alpha,\gamma) = h_{\mathcal{T}} \sum_{\{i,j\}\in\mathcal{E}} \frac{|v_i - v_j|^\alpha}{d_{ij}^\alpha} \sum_{K\in\mathcal{T}_{ij}} |K| \sum_{k\in\mathcal{N}_K} \frac{\psi'(v_k)}{(1-|v_k|)^\gamma}, \tag{3.25}$$

then

$$\lim_{m\to\infty} A^{(m)}(\alpha,\gamma) = 0. \tag{3.26}$$

Proof. Applying Hölder's inequality with exponents $\frac{p}{\alpha}$ and $\frac{p}{p-\alpha}$ where $p > 3 \geq \alpha$ we get

$$A^{(m)}(\alpha,\gamma) \leq h_{\mathcal{T}} \left(A_1^{(m)}\right)^{\alpha/p} \left(A_2^{(m)}\right)^{(p-\alpha)/p},$$

with

$$A_1^{(m)} = \sum_{\{i,j\}\in\mathcal{E}} \frac{|v_i - v_j|^p}{d_{ij}^p} \sum_{K\in\mathcal{T}_{ij}} |K| \sum_{k\in\mathcal{N}_K} \psi'(v_k),$$

and

$$A_2^{(m)} = \sum_{\{i,j\}\in\mathcal{E}} \sum_{K\in\mathcal{T}_{ij}} |K| \sum_{k\in\mathcal{N}_K} \frac{\psi'(v_k)}{(1-|v_k|)^{\frac{p\gamma}{p-\alpha}}}.$$

Let us reorder the sum in $A_2^{(m)}$. We have, from the definition of the dual mesh,

$$A_2^{(m)} = \frac{d(d+1)}{2} \sum_{k\in\mathcal{N}} \frac{\psi'(v_k)}{(1-|v_k|)^{\frac{p\gamma}{p-\alpha}}} \sum_{K\in\mathcal{T}_k} |K| = \frac{d(d+1)^2}{2} \left\| \Pi_{\mathcal{M}} \frac{\psi'(v)}{(1-|v|)^{\frac{p\gamma}{p-\alpha}}} \right\|_1.$$

Applying (3.14) in Lemma 3.10, we get

$$A_2^{(m)} \leq \frac{d(d+1)^2}{2} C_4^{\left(1,\frac{p\gamma}{p-\alpha}\right)}.$$

Therefore $A_2^{(m)}$ remains bounded for all $m \geq 0$. Since, from Lemma 3.17, we have that

$$A_1^{(m)} \leq \frac{\|f\|_{M(\Omega)}}{a_0 h_{\mathcal{T}}},$$

we conclude using the fact $p > 3 \geq \alpha$ that

$$A^{(m)}(\alpha,\gamma) \leq (h_{\mathcal{T}})^{1-\frac{\alpha}{p}} \left(\frac{\|f\|_{M(\Omega)}}{a_0}\right)^{\alpha/p} \left(A_2^{(m)}\right)^{(p-\alpha)/p},$$

which implies that (3.26) holds. □

We now turn to the proof of convergence to the entropy solution.

Proof of Theorem 3.9. Let us prove that $\overline{u}$ satisfies (3.2), for any function $T \in \mathcal{F}$ and $\phi \in C_c^\infty(\Omega)$ (see Rem. 3.2). Denoting by $s_T > 0$ such that $T'(s) = 0$ for all $|s| \geq s_T$, let us select $q \in (0,1)$ such that $\psi(q) > s_T + \max_{x\in\Omega} |\phi(x)|$.

Let us denote $\phi_i = \phi(z_i)$ for any $i \in \mathcal{N}$. We remark that

$$T(u_i - \phi_i) = T(\sigma_q(u_i) - \phi_i) = T(\psi_q(v_i) - \phi_i).$$

Letting $w = T(\psi_q(v) - P_{\mathcal{T}}\phi)$ in (3.8), we get

$$\sum_{\{i,j\}\in\mathcal{E}} (\Lambda_{ij}\Psi_{ij}(v_i,v_j) + S_{ij}(v))(v_i - v_j)(T(\psi_q(v_i) - \phi_i) - T(\psi_q(v_j) - \phi_j)) = \int_\Omega f\Pi_{\mathcal{T}} T(\psi_q(v) - P_{\mathcal{T}}\phi)\,\mathrm{d}x.$$

This implies

$$\sum_{\{i,j\}\in\mathcal{E}} (\Lambda_{ij}\Psi_{ij}(v_i,v_j) + S_{ij}(v))(v_i - v_j)T'(w_{ij})(\psi_q(v_i) - \psi_q(v_j) - (\phi_i - \phi_j)) = \int_\Omega f\Pi_{\mathcal{T}} T(\psi_q(v) - P_{\mathcal{T}}\phi)\,\mathrm{d}x,$$

where $w_{ij} \in I(\psi_q(v_i) - \phi_i, \psi_q(v_j) - \phi_j)$. We then have

$$A_{11}^{(m)} + A_{12}^{(m)} - A_{21}^{(m)} - A_{22}^{(m)} = A_3^{(m)}, \tag{3.27}$$

with

$$\begin{aligned}
A_{11}^{(m)} &:= \sum_{\{i,j\}\in\mathcal{E}} \Lambda_{ij}\Psi_{ij}(v_i,v_j)(v_i - v_j)T'(w_{ij})(\psi_q(v_i) - \psi_q(v_j)),\\
A_{12}^{(m)} &:= \sum_{\{i,j\}\in\mathcal{E}} S_{ij}(v)(v_i - v_j)T'(w_{ij})(\psi_q(v_i) - \psi_q(v_j)),\\
A_{21}^{(m)} &:= \sum_{\{i,j\}\in\mathcal{E}} \Lambda_{ij}\Psi_{ij}(v_i,v_j)(v_i - v_j)T'(w_{ij})(\phi_i - \phi_j)\\
A_{22}^{(m)} &:= \sum_{\{i,j\}\in\mathcal{E}} S_{ij}(v)(v_i - v_j)T'(w_{ij})(\phi_i - \phi_j),
\end{aligned}$$

and

$$A_3^{(m)} := \int_\Omega f\Pi_{\mathcal{T}} T(\psi_q(v) - P_{\mathcal{T}}\phi)\,\mathrm{d}x.$$

The remaining of the proof consists in studying the limit or the limitinf of each of these terms.

Term A_{11}

Using Lemma A.4 and $T' \geq 0$, we have

$$A_{11}^{(m)} \geq \sum_{\{i,j\}\in\mathcal{E}} \Lambda_{ij}T'(w_{ij})\Big(\widetilde{\psi}_q(v_i) - \widetilde{\psi}_q(v_j)\Big)^2 = A_{111}^{(m)} + A_{112}^{(m)}, \tag{3.28}$$

with, denoting by

$$w^K = \psi_q(v_{i_0}) - \phi_{i_0} \text{ with } i_0 \in \mathcal{N}_K \text{ such that } |\psi_q(v_{i_0}) - \phi_{i_0}| = \max_{i\in\mathcal{N}_K} |\psi_q(v_i) - \phi_i|, \tag{3.29}$$

and by $w_{\mathcal{T}}$ the function defined on Ω, equal *a.e.* to w^K on $K \in \mathcal{T}$, we have

$$A_{111}^{(m)} := \sum_{K\in\mathcal{T}} \sum_{\{i,j\}\in\mathcal{E}_K} \Lambda_{ij}^K T'\big(w^K\big)\Big(\widetilde{\psi}_q(v_i) - \widetilde{\psi}_q(v_j)\Big)^2 = \int_\Omega T'(w_{\mathcal{T}})\Lambda\nabla\Pi_{\mathcal{T}}\widetilde{\sigma}_q(u)\cdot\nabla\Pi_{\mathcal{T}}\widetilde{\sigma}_q(u)\,\mathrm{d}x,$$

and

$$A_{112}^{(m)} := \sum_{K\in\mathcal{T}} \sum_{\{i,j\}\in\mathcal{E}_K} \Lambda_{ij}^K \big(T'(w_{ij}) - T'\big(w^K\big)\big)\Big(\widetilde{\psi}_q(v_i) - \widetilde{\psi}_q(v_j)\Big)^2.$$

Definition (3.29) for w^K is motivated, on one hand, by defining a piecewise constant function on the elements, on the other hand, by the fact that $T'\big(w^K\big) \neq 0$ implies that $\max_{i\in\mathcal{N}_K}|\psi_q(v_i)| \le \psi(q)$, used in the proof of the convergence of the gradient. Note that, owing to the inequality $|b_i - b_{i_0}| \le \sum_{j\in\mathcal{N}_K}|b_j - b_{i_0}|$ for any $i, i_0 \in \mathcal{N}_K$ and $(b_i)_{i\in\mathcal{N}_K}$, we have the relation

$$\|T'(w_{\mathcal{T}}) - \Pi_{\mathcal{M}}T'(\psi_q(v) - P_{\mathcal{T}}\phi)\|_1 \quad\le\quad \max(|T''|)h_{\mathcal{T}} \sum_{K\in\mathcal{T}} \frac{|K|}{d+1} \sum_{i\in\mathcal{N}_K}\sum_{j\in\mathcal{N}_K} \frac{|\psi_q(v_i) - \psi_q(v_j)| + |\phi_i - \phi_j|}{d_{ij}}.$$

Remarking that $|\psi_q(v_i) - \psi_q(v_j)| \le \psi'(q)|v_i - v_j|$, we can apply the Cauchy–Schwarz inequality and the inequality

$$\sum_{K\in\mathcal{T}} \frac{|K|}{d+1} \sum_{i\in\mathcal{N}_K}\sum_{j\in\mathcal{N}_K} \frac{(v_i - v_j)^2}{d_{ij}^2} \le \sum_{K\in\mathcal{T}} \frac{|K|}{d+1} \sum_{i\in\mathcal{N}_K}\sum_{j\in\mathcal{N}_K} \frac{(\beta(v_i) - \beta(v_j))^2}{d_{ij}^2} = d\|\beta(v)\|_{1,2,\mathcal{M}}^2.$$

Owing to (A.2) in Lemma A.2 and to (3.13) in Lemma 3.10, this leads to

$$\lim_{m\to\infty}\left\|T'\Big(w_{\mathcal{T}}^{(m)}\Big) - \Pi_{\mathcal{M}^{(m)}}T'\Big(\psi_q\Big(v^{(m)}\Big) - P_{\mathcal{T}}^{(m)}\phi\Big)\right\|_1 = 0,$$

and therefore that

$$\lim_{m\to\infty}\left\|T'\Big(w_{\mathcal{T}}^{(m)}\Big) - T'(\sigma_q(\overline{u}) - \phi)\right\|_1 = 0.$$

Up to the extraction of a subsequence we can assume the convergence *a.e.* of $T'\Big(w_{\mathcal{T}}^{(m)}\Big)$ to $T'(\sigma_q(\overline{u}) - \phi)$.

Term A_{111}

We have $A_{111}^{(m)} = A_{1111}^{(m)} + 2A_{1112}^{(m)} - A_{1113}^{(m)}$ with

$$A_{1111}^{(m)} := \int_\Omega T'(w_{\mathcal{T}})\Lambda(\nabla\Pi_{\mathcal{T}}\widetilde{\sigma}_q(u) - \nabla\widetilde{\sigma}_q(\overline{u}))\cdot(\nabla\Pi_{\mathcal{T}}\widetilde{\sigma}_q(u) - \nabla\widetilde{\sigma}_q(\overline{u}))\,\mathrm{d}x \ge 0,$$

$$A_{1112}^{(m)} := \int_\Omega T'(w_{\mathcal{T}})\Lambda\nabla\Pi_{\mathcal{T}}\widetilde{\sigma}_q(u)\cdot\nabla\widetilde{\sigma}_q(\overline{u})\,\mathrm{d}x,$$

$$A_{1113}^{(m)} := \int_\Omega T'(w_{\mathcal{T}})\Lambda\nabla\widetilde{\sigma}_q(\overline{u})\cdot\nabla\widetilde{\sigma}_q(\overline{u})\,\mathrm{d}x.$$

The nonnegativity of $A_{1111}^{(m)}$ implies

$$\liminf_{m\to\infty} A_{1111}^{(m)} \ge 0.$$

By weak convergence in L^2 of $\nabla\Pi_{\mathcal{T}}\widetilde{\sigma}_q(u)$ (proved in Lem. 3.18) and strong convergence in L^2 of $T'(w_{\mathcal{T}})\nabla\widetilde{\sigma}_q(\overline{u})$ (indeed, $\|(T'(w_{\mathcal{T}}) - T'(\sigma_q(\overline{u}) - \phi))\nabla\widetilde{\sigma}_q(\overline{u})\|_2^2$ tends to 0 *a.e.* with being dominated by $2\max(T')|\nabla\widetilde{\sigma}_q(\overline{u})|^2 \in L^1(\Omega)$), we obtain

$$\lim_{m\to\infty} A_{1112}^{(m)} = \lim_{m\to\infty} A_{1113}^{(m)} = \int_\Omega T'(\sigma_q(\overline{u}) - \phi)\Lambda\nabla\widetilde{\sigma}_q(\overline{u})\cdot\nabla\widetilde{\sigma}_q(\overline{u})\,\mathrm{d}x.$$

Therefore

$$\liminf_{m\to\infty} A_{111}^{(m)} \ge \int_\Omega T'(\sigma_q(\overline{u}) - \phi)\Lambda\nabla\widetilde{\sigma}_q(\overline{u})\cdot\nabla\widetilde{\sigma}_q(\overline{u})\,\mathrm{d}x.$$

Term A_{112}

We remark that, for all $s \in (-1,1)$, since $\psi_q'(s) \le \psi'(q)$, the function ψ_q admits the Lipschitz constant $\psi'(q)$. Besides, we have, for all $a, b \in (-1,1)$,

$$\left(\widetilde{\psi}_q(a) - \widetilde{\psi}_q(b)\right)^2 = \left(\int_a^b \sqrt{\psi'(s)\psi_q'(s)}\,\mathrm{d}s\right)^2 \le \psi'(q)(\psi'(a)+\psi'(b))(a-b)^2.$$

This enables to write, denoting by C_T a bound of T'',

$$\left|A_{112}^{(m)}\right| \le C_T\psi'(q)\overline{\lambda}\theta_{\mathcal{T}}^2 \sum_{K\in\mathcal{T}} \sum_{\{k,l\}\in\mathcal{E}_K} (\psi'(q)|v_k - v_l| + C_\phi h_K) \sum_{\{i,j\}\in\mathcal{E}_K} \frac{|K|}{h_K^2}(v_i - v_j)^2(\psi'(v_i)+\psi'(v_j)).$$

We obtain, using $\psi'(v_i)+\psi'(v_j) \le \sum_{m\in\mathcal{N}_K} \psi'\big(v^{(m)}\big)$ and the Young inequality $|v_k - v_l|(v_i - v_j)^2 \le \frac{1}{3}|v_k - v_l|^3 + \frac{2}{3}|v_i - v_j|^3$, the inequality $\left|A_{112}^{(m)}\right| \le A_{1121}^{(m)} + A_{1122}^{(m)}$ with

$$A_{1121}^{(m)} := C_T\psi'(q)^2\overline{\lambda}\theta_{\mathcal{T}}^2 \sum_{K\in\mathcal{T}} \frac{|K|}{h_K^2} \sum_{\{k,l\}\in\mathcal{E}_K} \sum_{\{i,j\}\in\mathcal{E}_K} \left(\frac{1}{3}|v_k - v_l|^3 + \frac{2}{3}|v_i - v_j|^3\right) \sum_{\ell\in\mathcal{N}_K} \psi'(v_\ell),$$

and

$$A_{1122}^{(m)} := C_T\psi'(q)\overline{\lambda}\theta_{\mathcal{T}}^2 C_\phi h_{\mathcal{T}} \frac{d(d+1)}{2} \sum_{K\in\mathcal{T}} \sum_{\{i,j\}\in\mathcal{E}_K} \frac{|K|}{h_K^2}(v_i - v_j)^2(\psi'(v_i)+\psi'(v_j)).$$

We get, using $d_{ij} \le h_K$,

$$A_{1121}^{(m)} \le C_T\psi'(q)^2\overline{\lambda}\theta_{\mathcal{T}}^2 \frac{d(d+1)}{2} h_{\mathcal{T}} \sum_{K\in\mathcal{T}} |K| \sum_{\{i,j\}\in\mathcal{E}_K} \frac{|v_i - v_j|^3}{d_{ij}^3} \sum_{\ell\in\mathcal{N}_K} \psi'(v_\ell),$$

which provides

$$A_{1121}^{(m)} \le C_T\psi'(q)^2\overline{\lambda}\theta_{\mathcal{T}}^2 \frac{d(d+1)}{2} A^{(m)}(3,0),$$

where $A^{(m)}(\cdot,\cdot)$ is defined by (3.25), and

$$A_{1122}^{(m)} \le C_T\psi'(q)\overline{\lambda}\theta_{\mathcal{T}}^2 C_\phi \frac{d(d+1)}{2} A^{(m)}(2,0).$$

Applying Lemma 3.20, we get that

$$\lim_{m\to\infty} A_{1121}^{(m)} = \lim_{m\to\infty} A_{1122}^{(m)} = \lim_{m\to\infty} A_{112}^{(m)} = 0.$$

Hence the conclusion of the study of A_{11} is

$$\liminf_{m\to\infty} A_{11}^{(m)} \ge \int_\Omega T'(\sigma_q(\overline{u}) - \phi)\Lambda\nabla\widetilde{\sigma}_q(\overline{u}) \cdot \nabla\widetilde{\sigma}_q(\overline{u})\,\mathrm{d}x. \tag{3.30}$$

Term A_{12}

We have $A_{12}^{(m)} \ge 0$, therefore

$$\liminf_{m\to\infty} A_{12}^{(m)} \ge 0. \tag{3.31}$$

Term A_{21}

Using $v_{1,ij}$ defined by (3.23), $v_{2,ij}$ defined by (3.24), w^K and the function $w_{\mathcal{T}}$ defined above, we have $A_{21}^{(m)} = A_{211}^{(m)} + A_{212}^{(m)} + A_{213}^{(m)}$ with

$$A_{211}^{(m)} := \sum_{\{i,j\}\in\mathcal{E}} \Lambda_{ij}(u_i - u_j)T'\big(w^K\big)(\phi_i - \phi_j) = \int_\Omega T'(w_{\mathcal{T}})\Lambda\nabla\Pi_{\mathcal{T}}u \cdot \nabla\Pi_{\mathcal{T}}P_{\mathcal{T}}\phi\,\mathrm{d}x,$$
$$A_{212}^{(m)} := \sum_{\{i,j\}\in\mathcal{E}} \Lambda_{ij}(\psi'(v_{2,ij}) - \psi'(v_{1,ij}))(v_i - v_j)T'\big(w^K\big)(\phi_i - \phi_j),$$
$$A_{213}^{(m)} := \sum_{\{i,j\}\in\mathcal{E}} \Lambda_{ij}\Psi_{ij}(v_i, v_j)(v_i - v_j)\big(T'(w_{ij}) - T'\big(w^K\big)\big)(\phi_i - \phi_j).$$

We have, by weak convergence of $\nabla\Pi_{\mathcal{T}}u$ in L^r for any $1 < r < d/(d-1)$ and strong convergence of $T'(w_{\mathcal{T}})\nabla\Pi_{\mathcal{T}}P_{\mathcal{T}}\phi$ in $L^{r'}$ with $1/r + 1/r' = 1$ (recall that T' is bounded, hence dominated convergence applies),

$$\lim_{m\to\infty} A_{211}^{(m)} = \int_\Omega T'(\sigma_q(\overline{u}) - \phi)\Lambda\nabla\overline{u} \cdot \nabla\phi\,\mathrm{d}x.$$

Turning to $A_{212}^{(m)}$, we have, referring to the proof of Theorem 3.7,

$$A_{212}^{(m)} \le \max(T')B_{120}^{(m)},$$

which proves that this term tends to 0 as $m \to +\infty$.

Studying $A_{213}^{(m)}$, we write

$$\left|A_{213}^{(m)}\right| \le \max(T')\overline{\lambda}\sum_{K\in\mathcal{T}}\sum_{\{i,j\}\in\mathcal{E}_K}|K|\frac{\theta_{\mathcal{T}}^2}{h_K^2}\sum_{\ell\in\mathcal{N}_K}\psi'(v_\ell)|v_i - v_j|\sum_{\{k,l\}\in\mathcal{E}_K}(\psi'(q)|v_k - v_l| + C_\phi h_K)C_\phi h_K.$$

Following the treatment of $A_{112}^{(m)}$, where Young's inequality is replaced by $|v_i - v_j||v_k - v_l| \le \frac{1}{2}(v_i - v_j)^2 + \frac{1}{2}(v_k - v_l)^2$, we get

$$\left|A_{213}^{(m)}\right| \le \max(T')C_\phi\theta_{\mathcal{T}}^2\overline{\lambda}\frac{d(d+1)}{2}\Big(\psi'(q)A^{(m)}(2,0) + C_\phi A^{(m)}(1,0)\Big),$$

which also shows that this term tends to 0 as $m \to +\infty$.

Hence the conclusion of the study of A_{21} is

$$\lim_{m\to\infty} A_{21}^{(m)} = \int_\Omega T'(\sigma_q(\overline{u}) - \phi)\Lambda\nabla\overline{u} \cdot \nabla\phi\,\mathrm{d}x. \tag{3.32}$$

Term A_{22}

Comparing $A_{22}^{(m)}$ with Term $B_{20}^{(m)}$ in the proof of Theorem 3.7, we remark that

$$\left|A_{22}^{(m)}\right| \le \max(T')B_{20}^{(m)},$$

which shows that

$$\lim_{m\to\infty} A_{22}^{(m)} = 0. \tag{3.33}$$

Term A_3

We have, by almost everywhere and dominated convergence,

$$\lim_{m\to\infty} A_3^{(m)} = \int_\Omega T(\sigma_q(\overline{u}) - \phi)f\,\mathrm{d}x. \tag{3.34}$$

Conclusion from (3.27), (3.30)–(3.34)

We deduce from these equations that

$$\int_\Omega T'(\sigma_q(\overline{u})-\phi)\Lambda\nabla\widetilde{\sigma}_q(\overline{u})\cdot\nabla\widetilde{\sigma}_q(\overline{u})-\int_\Omega T'(\sigma_q(\overline{u})-\phi)\Lambda\nabla\overline{u}\cdot\nabla\phi\,\mathrm{d}x\le\int_\Omega T(\sigma_q(\overline{u})-\phi)f(x)\,\mathrm{d}x. \tag{3.35}$$

Since $\psi(q)>s_T+\max|\phi|$, recalling that $\sigma_q=\widetilde{\sigma}_q=\mathrm{Id}$ on $[-\psi(q),\psi(q)]$, we get that $T'(\sigma_q(\overline{u})-\phi)=T'(\overline{u}-\phi)$, $\widetilde{\sigma}_q'(\overline{u})=1$ if $T'(\sigma_q(\overline{u})-\phi)\neq 0$, and $T(\sigma_q(\overline{u})-\phi)=T(\overline{u}-\phi)$. This allows to conclude that (3.2) holds, which implies that $\overline{u}$ is the unique entropy solution of the problem. From the uniqueness property of the limit, we deduce that the convergence properties proved by Lemma 3.18 (except the almost everywhere convergence) hold for the whole sequence.

Let us now turn to the strong convergence of the gradient

We let $\phi=0$, and, for a given function $T\in\mathcal{F}$, we again denote by $s_T>0$ such that $T'(s)=0$ for all $|s|\ge s_T$, and define $q\in(0,1)$ such that $\psi(q)>s_T$. Letting $w=T(\psi_q(v))$ in (3.8), we consider all the terms A_i, $i=1,11,2,\dots$ as defined above, and we get (3.27), which leads, owing to (3.28) to

$$A_{111}^{(m)}+A_{112}^{(m)}+A_{12}^{(m)}-A_{21}^{(m)}-A_{22}^{(m)}\le A_3^{(m)}.$$

Since $\phi=0$ implies $A_{21}^{(m)}=0$ and using $A_{12}^{(m)}\ge 0$, we obtain

$$A_{111}^{(m)}+A_{112}^{(m)}-A_{22}^{(m)}\le A_3^{(m)}.$$

Since we prove above that the terms $A_{112}^{(m)}$ and $A_{22}^{(m)}$ tend to 0 as $m\to\infty$, using Lemma A.6, we get

$$\limsup_{m\to\infty}A_{111}^{(m)}\le\lim_{m\to\infty}A_3^{(m)}=\int_\Omega T(\sigma_q(\overline{u}))f\,\mathrm{d}x=\int_\Omega T(\overline{u})f\,\mathrm{d}x=\int_\Omega T'(\overline{u})\Lambda\nabla\overline{u}\cdot\nabla\overline{u}.$$

We have $A_{1111}^{(m)}=A_{111}^{(m)}-2A_{1112}^{(m)}+A_{1113}^{(m)}$, which leads to

$$\limsup_{m\to\infty}A_{1111}^{(m)}\le\limsup_{m\to\infty}A_{111}^{(m)}-2\int_\Omega T'(\overline{u})\Lambda\nabla\overline{u}\cdot\nabla\overline{u}+\int_\Omega T'(\overline{u})\Lambda\nabla\overline{u}\cdot\nabla\overline{u}\le 0.$$

This proves that

$$\lim_{m\to\infty}\int_\Omega T'(w_{\mathcal{T}^{(m)}})\Lambda\Big(\nabla\Pi_{\mathcal{T}^{(m)}}\widetilde{\sigma}_q\Big(u^{(m)}\Big)-\nabla\widetilde{\sigma}_q(\overline{u})\Big)\cdot\Big(\nabla\Pi_{\mathcal{T}^{(m)}}\widetilde{\sigma}_q\Big(u^{(m)}\Big)-\nabla\widetilde{\sigma}_q(\overline{u})\Big)\,\mathrm{d}x=0,$$

and therefore that

$$\lim_{m\to\infty}\int_\Omega T'(w_{\mathcal{T}^{(m)}})\Big|\nabla\Pi_{\mathcal{T}^{(m)}}\widetilde{\sigma}_q\Big(u^{(m)}\Big)-\nabla\widetilde{\sigma}_q(\overline{u})\Big|^2\,\mathrm{d}x=0.$$

From Definition (3.29) of $w_{\mathcal{T}}$, $T'(w_{\mathcal{T}^{(m)}}(x))\neq 0$ means that, if $K\in\mathcal{T}^{(m)}$ is such that $x\in K$, we have $\max_{i\in\mathcal{N}_K}|\psi_q(v_i)|\le\psi(q)$, which implies that $\nabla\Pi_{\mathcal{T}^{(m)}}\widetilde{\sigma}_q\big(u^{(m)}\big)(x)=\nabla\Pi_{\mathcal{T}^{(m)}}u^{(m)}(x)$. We thus get

$$\lim_{m\to\infty}\int_\Omega T'(w_{\mathcal{T}^{(m)}})\Big|\nabla\Pi_{\mathcal{T}^{(m)}}u^{(m)}-\nabla\widetilde{\sigma}_q(\overline{u})\Big|^2\,\mathrm{d}x=0. \tag{3.36}$$

The remaining of the proof is dedicated to show that (3.36) implies the strong convergence of $\nabla\Pi_{\mathcal{T}^{(m)}}u^{(m)}$ to $\nabla\overline{u}$.

We now denote a given representative of the functions $\overline{u}$, $\nabla\overline{u}$ and for any $m\in\mathbb{N}$, of $w_{\mathcal{T}^{(m)}}$ and of $\nabla\Pi_{\mathcal{T}^{(m)}}u^{(m)}$, defined everywhere in Ω, by the same notation.

Step 1. Construction of a decreasing sequence $(N_n)_{n\in\mathbb{N}}$ of infinite subsets of $\mathbb{N}$, and of a sequence $(\Omega_n)_{n\in\mathbb{N}}$ of subsets of Ω such that

- $|\Omega\setminus\Omega_n|\to 0$ as n tends to infinity.
- For all $n\geq 1$ and $x\in\Omega_n$, $\nabla\Pi_{\mathcal{T}^{(m)}}u^{(m)}(x)$ converges to $\nabla\overline{u}(x)$ as $m\in N_n$ tends to infinity.

We let $N_0=\mathbb{N}$.

We suppose that, for a given strictly positive integer n, is given an infinite set $N_{n-1}\subset\mathbb{N}$.

We then define $T_n\in\mathcal{F}$ such that $T_n(s)=s$ for all $|s|\leq n$, and $q\in(0,1)$ such that $n\leq s_{T_n}\leq\psi(q)$ (which means that $T_n'(s)=0$ for all $|s|\geq\psi(q)$). Letting $\Omega_{0,n}$ be defined by

$$\Omega_{0,n}:=\{x\in\Omega;|\overline{u}(x)|\leq n\},$$

we get $|\Omega\setminus\Omega_{0,n}|\leq\frac{\|\overline{u}\|_1}{n}$. We then have $\widetilde{\sigma}_q(\overline{u}(x))=\overline{u}(x)$ and $T_n'(\widetilde{\sigma}_q(\overline{u}(x)))=T_n'(\overline{u}(x))=1$ for all $x\in\Omega_{0,n}$.

We now define $\Omega_n\subset\Omega_{0,n}$ with $|\Omega_{0,n}\setminus\Omega_n|=0$ and an infinite set $N_n\subset N_{n-1}$ such that

- for any $x\in\Omega_n$, we have $\nabla\overline{u}(x)=\nabla\widetilde{\sigma}_q(\overline{u})(x)$, since it holds $\nabla\widetilde{\sigma}_q(\overline{u})(x)=\widetilde{\sigma}_q'(\overline{u}(x))\nabla\overline{u}(x)$ for *a.e.* $x\in\Omega_{0,n}$ from [25];
- using the convergence in L^1 of $T_n'(w_{\mathcal{T}^{(m)}})$ to $T_n'(\overline{u})$ and extracting a subsequence, for any $x\in\Omega_n$, $T_n'(w_{\mathcal{T}^{(m)}}(x))$ converges to $T_n'(\overline{u}(x))=1$ as $m\in N_n$ tends to infinity;
- using (3.36) and extracting a subsequence, for any $x\in\Omega_n$, $T_n'(w_{\mathcal{T}^{(m)}}(x))\left|\nabla\Pi_{\mathcal{T}^{(m)}}u^{(m)}(x)-\nabla\widetilde{\sigma}_q(\overline{u}(x))\right|^2$ tends to 0 as $m\in N_n$ tends to infinity.

We have, for any $x\in\Omega_n$, that $T_n'(w_{\mathcal{T}^{(m)}}(x))>\frac{1}{2}$ for $m\in N_n$ large enough, which means that $\nabla\Pi_{\mathcal{T}^{(m)}}u^{(m)}(x)$ tends to $\nabla\widetilde{\sigma}_q(\overline{u}(x))=\nabla\overline{u}(x)$.

We then have $|\Omega\setminus\Omega_n|=|\Omega\setminus\Omega_{0,n}|\leq\frac{\|\overline{u}\|_1}{n}$.

Step 2. Diagonal process and convergence almost everywhere of the gradient.

Since all the infinite sets $(N_n)_{n\in\mathbb{N}}$ are ordered, we denote by $N_n^{(k)}$ the kth element of N_n. The property $N_n\subset N_{n-1}$ implies that $N_n^{(n)}>N_{n-1}^{(n-1)}$. The diagonal process consists in defining

$$N_\infty=\left\{N_n^{(n)},n\in\mathbb{N}\right\},$$

which therefore satisfies that $\left\{N_\infty^{(k)},\ k\geq n\right\}\subset N_n$. Then we denote by

$$\Omega_\infty=\bigcup_{n\geq 1}\Omega_n.$$

We remark that $|\Omega\setminus\Omega_\infty|\leq\frac{\|\overline{u}\|_1}{n}$ for all $n\geq 1$ implies that $|\Omega\setminus\Omega_\infty|=0$. For any $x\in\Omega_\infty$, there exists $n\geq 1$ such that $x\in\Omega_n$. Since $\left\{N_\infty^{(k)},\ k\geq n\right\}\subset N_n$, we deduce that $\nabla\Pi_{\mathcal{T}^{(m)}}u^{(m)}(x)$ converges to $\nabla\overline{u}(x)$ as $m\in N_\infty$ tends to $+\infty$.

This concludes the proof that there exists a subset $\Omega_\infty\subset\Omega$ such that $|\Omega\setminus\Omega_\infty|=0$, and a subsequence of approximate solutions indexed by $m\in N_\infty$, such that, for all $x\in\Omega_\infty$, $\nabla\Pi_{\mathcal{T}^{(m)}}u^{(m)}(x)$ converges to $\nabla\overline{u}(x)$ as $m\in N_\infty$ tends to $+\infty$.

Step 3. Convergence in L^q for $1<q<d/(d-1)$.

We now apply a classical reasoning. Since $\left(\nabla\Pi_{\mathcal{T}^{(m)}}u^{(m)}\right)_{m\in N_\infty}$ is bounded in $L^r(\Omega)$ for a given $r\in(1,d/(d-1))$ by Lemma 3.14, it is therefore equi-integrable. Since Ω is bounded, and since $\left(\nabla\Pi_{\mathcal{T}^{(m)}}u^{(m)}\right)_{m\in N_\infty}$ converges almost everywhere to $\nabla\overline{u}$, we deduce from Vitali's theorem that $\left(\nabla\Pi_{\mathcal{T}^{(m)}}u^{(m)}\right)_{m\in N_\infty}$ converges in $L^1(\Omega)$ to $\nabla\overline{u}$. By interpolation L^q-L^r, we get that this convergence holds for any $q\in(1,d/(d-1))$. By uniqueness of the limit, we conclude that the whole sequence converges for this topology. □

4. Numerical tests

We present here some illustrations of the behavior of the numerical schemes (2.6) and (3.8) with $d = 2$ or $d = 3$. The implementation of Scheme (3.8) is simply done in the case $a_0 = 0$, using a Picard iteration for handling the term $\Psi_{ij}(v_i, v_j)$.

Let us detail the implementation of Scheme (2.6). We also adopt a simple Picard iteration method for approximating the solution of the resulting systems of nonlinear equations, since we observe good convergence properties in the studied test cases. It consists in computing the sequence $\left(v_{\mathcal{T}}^{(k)}\right)_{k=0,\dots,N}$ such that $v_{\mathcal{T}}^{(0)} = 0$ and, for all $k = 1, \dots, N$,

$$\int_\Omega \psi'\left(v_{\mathcal{T}}^{(k-1)}\right) \Lambda_{\mathcal{T}} \nabla v_{\mathcal{T}}^{(k)} \cdot \nabla w \, \mathrm{d}x = \int_\Omega f w \, \mathrm{d}x \text{ for all } w \in \mathcal{V}_{\mathcal{T}}.$$

The value N is determined by the criterion $\|v_{\mathcal{T}}^{(k)} - v_{\mathcal{T}}^{(k-1)}\|_{L^\infty(\Omega)} \le \sigma_{\mathrm{tol}}$. In the following examples, we let $\sigma_{\mathrm{tol}} = 10^{-8}$ and N is about 10 in the numerical examples. This choice of an L^∞ criterion is done in order to accurately approximate the large values of the unknown function; it can be expected that an L^1 criterion would lead to a smaller number of iterations.

The implementation of this Picard iteration method leads to the evaluation of $\int_K \psi'\left(v_{\mathcal{T}}^{(k-1)}\right) \mathrm{d}x$ for any $K \in \mathcal{T}$. We use the formulas given in Section A.1, which implies to numerically determine which case of equality is satisfied by the values $(v_i)_{i=1,\dots,d+1}$. This is done by comparing the differences between the values with 10^{-3} (smaller values lead to less precise results due to the divisions by these differences).

A second problem in the implementation of the method, in the case where the measure f is in fact an element of $L^1(\Omega)$ and $f(x)$ is singular at some points x of the domain, is the computation of the right-hand sides $\int_\Omega f(x)\varphi_i(x) \, \mathrm{d}x$. The key point for the precision of the method is to preserve the exactness of the integrals: the use of approximate quadrature formulas with Gauss points leads to very poor accuracy in this case. We complete this goal by replacing $\int_\Omega f(x)\varphi_i(x) \, \mathrm{d}x$ with $\int_{\omega_i} f(x) \, \mathrm{d}x$, where ω_i is a dual cell associated with the vertex z_i (the exact shape of ω_i has no influence on the precision of the computation), and by computing the exact integration of $\int_{\omega_i} f(x) \, \mathrm{d}x$.

Finally, let us observe that, in the examples below, we compare numerical solutions computed with polygonal meshes to analytical solutions available on non-polygonal domains (circles, cylinder). These analytical solutions are vanishing at the boundary of these domains, which leads to an error lower than h^2, where h is the size of the mesh.

4.1. Case $d = 2$, measure

We consider the case where $\Omega = B(0, 1)$ and

$$\Lambda(x) = \begin{pmatrix} 1 + (\beta - 1)\frac{x_1^2}{x_1^2 + x_2^2} & (\beta - 1)\frac{x_1 x_2}{x_1^2 + x_2^2} \\ (\beta - 1)\frac{x_1 x_2}{x_1^2 + x_2^2} & 1 + (\beta - 1)\frac{x_2^2}{x_1^2 + x_2^2} \end{pmatrix} \text{ with } \beta = 5 \text{ for all } x \in \Omega. \tag{4.1}$$

This heterogeneous and anisotropic diffusion field corresponds in 2D to the case described by Prignet [23] and Serrin [24]. We replace the average values of Λ in the elements by the values at the center of gravity of the elements.

We let $f = \beta\delta_{(0,0)}$, which corresponds to the analytical solution given by $\overline{u}(x) = -\frac{1}{2\pi} \log|x|$ (see the right part of Fig. 4 for a representation of the numerical solution).

In this test case, the solution is no longer in $H_0^1(\Omega)$ nor in $L^\infty(\Omega)$.

We use triangular meshes which are refined around the point $(0, 0)$ (see the left part of Fig. 4), and we compute $\psi(v_{\mathcal{T}})$ solution to Scheme (2.6), $\Pi_{\mathcal{T}} u$ solution to the control-volume finite-element scheme (3.8) with $a_0 = 0$ (as detailed in Sect. 3, this value suffices for the convergence of the scheme to a weak solution, but

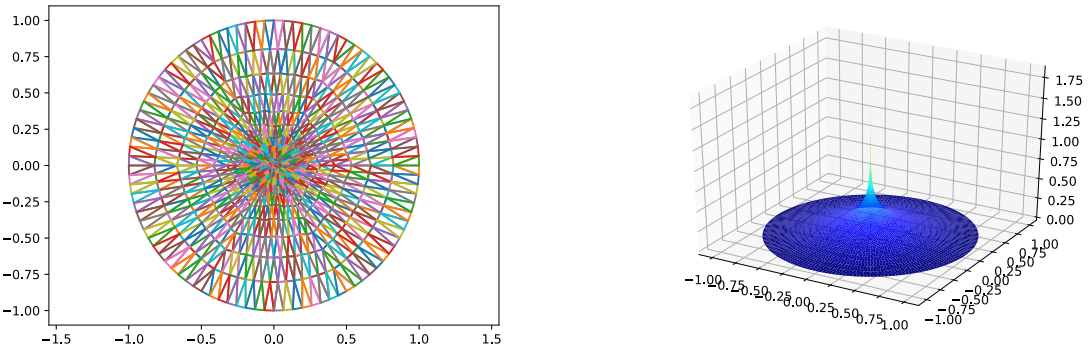

FIGURE 4. Domain Ω and mesh in the case $n = 817$ (*left*), numerical solution with $n = 12\,481$ (*right*).

TABLE 1. $L^1(\Omega)$ errors in the measure 2D case (n is the total number of vertices).

n	$\mathrm{err}_{u,\mathrm{I}}$	Order	$\mathrm{err}_{u,\mathrm{II}}$	Order	$\mathrm{err}_{u,\mathrm{III}}$	Order	$\mathrm{err}_{\nabla u,\mathrm{I}}$	Order	$\mathrm{err}_{\nabla u,\mathrm{II}}$	Order	$\mathrm{err}_{\nabla u,\mathrm{III}}$	Order
217	0.00437		0.0151		0.00523		0.193		0.198		0.221	
817	0.00106	2.14	0.00635	1.31	0.00128	2.12	0.101	0.98	0.106	0.94	0.116	0.97
3169	0.000263	2.06	0.00289	1.16	0.000318	2.05	0.0513	1.00	0.0542	0.99	0.0590	1.00
12 481	0.0000649	2.04	0.00138	1.08	0.0000793	2.03	0.0258	1.00	0.0274	1.00	0.0297	1.00
49 537	0.0000162	2.01	0.000672	1.04	0.0000198	2.01	0.0129	1.01	0.0138	1.00	0.0149	1.00

not to the entropy solution; in practice, increasing a_0 leads to an increase of the observed numerical error) and $u_{\mathcal{T}} \in \mathcal{V}_{\mathcal{T}}$ solution to the linear scheme (1.6). These solutions are respectively denoted by I, II and III in the tables below.

We let $\alpha = 10$ in the definition (2.3) of ψ_α; in this case, the Picard iterations do not converge with $\alpha = 1$. The quantities $\mathrm{err}_{u,\mathrm{I}} := \|\psi_\alpha(v_{\mathcal{T}}) - \overline{u}\|_{L^1(\Omega)}$ for Scheme I, $\mathrm{err}_{u,\mathrm{II}} := \|\Pi_{\mathcal{T}}\psi_\alpha(v_{\mathcal{T}}) - \overline{u}\|_{L^1(\Omega)}$ for Scheme II, and $\mathrm{err}_{u,\mathrm{III}} := \|u_{\mathcal{T}} - \overline{u}\|_{L^1(\Omega)}$ for Scheme (1.6), $\mathrm{err}_{\nabla u,\mathrm{I}} := \|\nabla\psi_\alpha(v_{\mathcal{T}}) - \nabla\overline{u}\|_{L^1(\Omega)}$ for Scheme I, $\mathrm{err}_{\nabla u,\mathrm{II}} := \|\nabla\Pi_{\mathcal{T}}\psi_\alpha(v_{\mathcal{T}}) - \nabla\overline{u}\|_{L^1(\Omega)}$ for Scheme II, and $\mathrm{err}_{\nabla u,\mathrm{III}} := \|\nabla u_{\mathcal{T}} - \nabla\overline{u}\|_{L^1(\Omega)}$ for Scheme (1.6) are provided in Table 1, using 5 meshes whose the total number of vertices is denoted by n. In this table, we define the meshsize as $n^{-1/2}$ and we compute the order of convergence with respect to the preceding line. The errors on the gradients are computed at the mid-edges of the triangles.

We observe that the non-linear method I provides slightly more accurate results than the linear method III, and that a numerical order, approximately equal to 2, is observed for $\overline{u}$ and an order 1 is observed for $\nabla\overline{u}$. Method II seems to provide an order 1, both for $\overline{u}$ and $\nabla\overline{u}$.

4.2. Case $d = 2$, $f \in L^1(\Omega)$

We consider the case where $\Omega = B\big(0, \frac{1}{2}\big)$ and Λ is defined by (4.1). We let $f(x) = \beta\gamma(1-\gamma)(-\log|x|)^{\gamma-2}/|x|^2$ with $\gamma = 3/4$, which corresponds to the analytical solution given by $\overline{u}(x) = (-\log|x|)^\gamma - (-\log(\frac{1}{2}))^\gamma$ (see Fig. 5 for a representation of the numerical solution).

In this test case, the solution is no longer in $H_0^1(\Omega)$ nor in $L^\infty(\Omega)$ (recall that it belongs to $H_0^1(\Omega)$ only for $\gamma \in \big(0, \frac{1}{2}\big)$). The right-hand side f is then in $L^1(\Omega)$, but this is not the case for the product $f\overline{u}$ which is not locally integrable around the point 0 (this prevents from using the solution as test function at the continuous level).

We use triangular meshes which are refined around the point $(0,0)$ (see Fig. 5 for an example of solution computed with such a mesh), and we again compute the solutions to Schemes I given by (2.6), II given by (3.8)

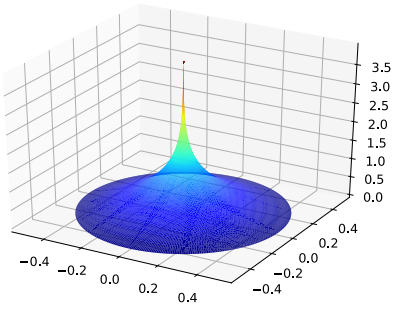

FIGURE 5. Numerical solution with $n = 12\,481$.

TABLE 2. $L^1(\Omega)$ errors in the 2D case (n is the total number of vertices).

n	$\mathrm{err}_{u,\mathrm{I}}$	Order	$\mathrm{err}_{u,\mathrm{II}}$	Order	$\mathrm{err}_{u,\mathrm{III}}$	Order	$\mathrm{err}_{\nabla u,\mathrm{I}}$	Order	$\mathrm{err}_{\nabla u,\mathrm{II}}$	Order	$\mathrm{err}_{\nabla u,\mathrm{III}}$	Order
217	0.00272		0.0735		0.0136		0.335		0.667		0.584	
817	0.000896	1.68	0.03854	0.97	0.00315	2.21	0.177	0.96	0.386	0.83	0.319	0.91
3169	0.000308	1.58	0.0203	0.95	0.000770	2.08	0.0896	1.00	0.207	0.92	0.163	0.99
12 481	0.0000840	1.90	0.0105	0.96	0.000197	1.99	0.0446	1.02	0.107	0.96	0.0817	1.01
49 537	0.0000162	2.39	0.00533	0.98	0.0000485	2.03	0.0223	1.01	0.0546	0.98	0.0409	1.00

with $a_0 = 0$ and III given by (1.6). Letting $\alpha = 10$ and using the same notations as in the preceding section, we obtain the results provided in Table 2.

The orders of convergence are similar to the ones observed in Table 1.

4.3. Case $d = 3$

In accordance with the introduction of this paper, we detail the example provided by Prignet [23] for the non-uniqueness of a solution in the sense of Definition 1.1. One considers $\Omega = B\big(0, \frac{1}{2}\big) \times (-1, 1) \subset \mathbb{R}^3$, and one defines the following diffusion field: denoting any point $x \in \Omega$ with $x = (x_1, x_2, x_3)$, let us define

$$\Lambda(x) = \begin{pmatrix} 1 + (\beta - 1)\frac{x_1^2}{x_1^2 + x_2^2} & (\beta - 1)\frac{x_1 x_2}{x_1^2 + x_2^2} & 0 \\ (\beta - 1)\frac{x_1 x_2}{x_1^2 + x_2^2} & 1 + (\beta - 1)\frac{x_2^2}{x_1^2 + x_2^2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.2}$$

with $\beta = 16$ (we then have $\beta = \frac{1}{\varepsilon^2}$ with $\varepsilon = \frac{1}{4} \in (0, \frac{1}{3})$). Then, for any pair of reals $(\eta_1, \eta_2) \neq (0, 0)$, the function

$$\overline{w}_{\eta_1,\eta_2}(x) = (\eta_1 x_1 + \eta_2 x_2)\left(\sqrt{x_1^2 + x_2^2}\right)^{1-d-\varepsilon} \quad \text{for any } x \in \Omega$$

is shown to satisfy:

- $-\mathrm{div}(\Lambda(x)\nabla \overline{w}_{\eta_1,\eta_2})(x) = 0$ *a.e.* in Ω,
- $\overline{w}_{\eta_1,\eta_2} \in W^{1,r}(\Omega)$ for $r \in \left(1, \frac{d}{d-1}\right)$ and $\overline{w}_{\eta_1,\eta_2} \notin H^1(\Omega)$,
- the restriction of $\overline{w}_{\eta_1,\eta_2}$ to the boundary of Ω belongs to $H^{1/2}(\partial\Omega)$, which means that there exists $\overline{z}_{\eta_1,\eta_2} \in H^1(\Omega)$, with the same trace, solution to the non-homogeneous Dirichlet problem

$$\forall w \in H_0^1(\Omega), \quad \int_\Omega \Lambda \nabla \overline{z}_{\eta_1,\eta_2} \cdot \nabla w \, \mathrm{d}x = 0.$$

TABLE 3. $L^1(\Omega)$ errors in the 3D case (n is the total number of vertices).

n	$\mathrm{err}_{u,\mathrm{I}}$	Order	$\mathrm{err}_{u,\mathrm{II}}$	Order	$\mathrm{err}_{u,\mathrm{III}}$	Order	$\mathrm{err}_{\nabla u,\mathrm{I}}$	Order	$\mathrm{err}_{\nabla u,\mathrm{II}}$	Order	$\mathrm{err}_{\nabla u,\mathrm{III}}$	Order
135	0.139		0.260		0.128		5.19		4.00		5.72	
765	0.0221	3.18	0.245	0.10	0.0661	1.14	2.64	1.17	3.28	0.34	3.56	0.82
5049	0.00911	1.41	0.200	0.32	0.0220	1.75	1.33	1.09	2.55	0.40	1.92	0.98
36 465	0.00282	1.78	0.151	0.43	0.00591	1.99	0.651	1.08	1.85	0.49	0.974	1.03
276 705	0.000731	2.00	–		0.00146	2.07	0.319	1.06	–		0.485	1.03

Therefore, problem (1.5) with $f = 0$ has both the strong solution 0 and the weak solutions $\overline{w}_{\eta_1,\eta_2} - \overline{z}_{\eta_1,\eta_2} \neq 0$ for any $(\eta_1, \eta_2) \in \mathbb{R}^2 \setminus \{(0,0)\}$.

We then denote, for all $x = (x_1, x_2, x_3)$, by $\overline{v}(x) = \left(-\log\left(\sqrt{x_1^2 + x_2^2}\right)\right)^\gamma - \left(-\log\left(\frac{1}{2}\right)\right)^\gamma$, with $\gamma = 3/4$, and by

$$g(x) = -\mathrm{div}(\Lambda \nabla \overline{v})(x) = \beta\gamma(1-\gamma)\frac{\left(-\log\left(\sqrt{x_1^2 + x_2^2}\right)\right)^{\gamma-2}}{x_1^2 + x_2^2},$$

and we define

$$f(x) = \left(1 - x_3^2\right) g(x) + 2\overline{v}(x),$$

for all $x \in \Omega$. Then the function given by $\overline{u}(x) = \left(1 - x_3^2\right)\overline{v}(x)$ is a weak solution in the sense of Definition 1.1, as well as $\overline{u} + \overline{w}_{\eta_1,\eta_2} - \overline{z}_{\eta_1,\eta_2}$ for any $(\eta_1, \eta_2) \in \mathbb{R}^2 \setminus \{(0,0)\}$. In this test case, none of these solutions belongs to $H_0^1(\Omega)$ or to $L^\infty(\Omega)$. We know from Prignet [23] that, for $(\eta_1, \eta_2) \neq (0,0)$, the truncated functions $T_k \overline{w}_{\eta_1,\eta_2}$, for any $k > 0$, are not in $H^1(\Omega)$. This is not the case for $\overline{u}$. We extend the 2D refined triangular meshes of $B\left(0, \frac{1}{2}\right)$ (see the preceding section) on Ω by generating prisms with constant height, which are then shared into 3 tetrahedra. Letting $\alpha = 1$, the following $L^1(\Omega)$ errors compared to that given by the linear method are provided in Table 3, with respect to the weak solution $\overline{u}$ (the references to the 3 schemes compared in this numerical section being the same as in the 2D case). In this table, we define the meshsize as $n^{-1/3}$ for computing the order of convergence.

One notices that all schemes seem to converge to $\overline{u}$ (recall that we proved that the truncations of the limit belong to $H_0^1(\Omega)$, which excludes any solution including a non-zero term $\overline{w}_{\eta_1,\eta_2} - \overline{z}_{\eta_1,\eta_2}$). The non-linear method I seems to provide slightly more accurate results than the linear method III, with a numerical order of convergence close to 2 for the values and 1 for the gradients considering the finest meshes. The linear scheme was not expected in this case to numerically converge to any weak solution different from $\overline{u}$, since in the case where $f = 0$, the only solution of the linear scheme is 0. The numerical order of convergence for method II seems to be closer to 1/2 than to 1, both for $\overline{u}$ and $\nabla\overline{u}$.

5. Some open and closed problems

The two nonlinear methods presented in this paper are proved to converge to a weak solution of the linear elliptic problem with general measure data and heterogeneous anisotropic diffusion fields, which does not seem to have been proved for earlier schemes. The convergence of the nonlinear CVFE scheme to the entropy weak solution seems also to be the first one in this framework. But it was necessary to assess the numerical performance of these schemes, which is done in Section 4, where we present numerical results for 2D and 3D cases obtained with these two schemes and with the simpler linear scheme (1.6) (for which no convergence properties are proved for general meshes and diffusion field). We observe that the accuracy of the nonlinear finite element scheme is comparable to that of the simpler linear scheme (1.6), but that of the CVFE scheme is disappointing.

Two questions (at least) remain open problems.

The first one is to prove the convergence for a stronger topology of the gradient of the numerical solution obtained by the nonlinear finite element scheme, since it seems to be observed in the numerical results.

The second one is to prove that the nonlinear finite element scheme or the nonlinear CVFE scheme without the stabilisation term converge to the entropy solution, which could be expected since the entropy solution is in fact designed to be the limit of the solutions to regularised problems.

Appendix A. Technical lemmas

We first recall the following Sobolev inequalities (see [1]).

Lemma A.1. *Let Ω be an open bounded Lipschitz subset of $\mathbb{R}^d$ and let $r \in [1, +\infty)$. Let $q \in \left[1, \frac{rd}{d-r}\right]$ if $r < d$, $q \in [1, +\infty)$ if $r = d$ and $q \in [1, +\infty]$ if $r > d$. Then there exists $C_{\rm sob}^{(r,q)}$, also depending on d and $|\Omega|$, such that for any $u \in W_0^{1,r}(\Omega)$ we have*

$$\|u\|_q \le C_{\rm sob}^{(r,q)} \|\nabla u\|_r.$$

A.1. Computation of some integrals in simplices

We consider here the discrete framework of Section 2. Let us compute, in the cases $d = 2$ and $d = 3$, the quantity $\int_K \zeta(v_{\mathcal{T}})\,\mathrm{d}x$, where ζ is any continuous function on $\mathbb{R}$. For any continuous function μ on $\mathbb{R}$, we denote by $\mathcal{I}(\mu)$ the primitive of μ equal to 0 at point 0. For a given $K \in \mathcal{T}$, let us denote the values of $v_{\mathcal{T}}$ at the vertices of K by $(v_i)_{i=1,\dots,d+1}$. The following holds:

Case $d = 2$.

– If $v_1 = v_2 = v_3$, then

$$\frac{1}{|K|}\int_K \zeta(v_{\mathcal{T}}(x))\,\mathrm{d}x = \zeta(v_1).$$

– If $v_1 \neq v_2 = v_3$, then

$$\frac{1}{|K|}\int_K \zeta(v_{\mathcal{T}}(x))\,\mathrm{d}x = 2\left(\frac{\mathcal{I}(\zeta)(v_2)}{(v_2 - v_1)} - \frac{\mathcal{I}^2(\zeta)(v_2) - \mathcal{I}^2(\zeta)(v_1)}{(v_2 - v_1)^2}\right).$$

– In the case where all the values v_i are distinct, we get

$$\frac{1}{|K|}\int_K \zeta(v_{\mathcal{T}}(x))\,\mathrm{d}x = 2\sum_{i=1}^{3} \frac{\mathcal{I}^2(\zeta)(v_i)}{\prod_{j\in\{1,2,3\}\setminus\{i\}}(v_i - v_j)}.$$

Case $d = 3$.

– If $v_1 = v_2 = v_3 = v_4$, then

$$\frac{1}{|K|}\int_K \zeta(v_{\mathcal{T}}(x))\,\mathrm{d}x = \zeta(v_1).$$

– If $v_1 \neq v_2 = v_3 = v_4$, then

$$\frac{1}{|K|}\int_K \zeta(v_{\mathcal{T}}(x))\,\mathrm{d}x = 6\left(\frac{\mathcal{I}(\zeta)(v_2)}{2(v_2 - v_1)} - \frac{\mathcal{I}^2(\zeta)(v_2)}{(v_2 - v_1)^2} + \frac{\mathcal{I}^3(\zeta)(v_2) - \mathcal{I}^3(\zeta)(v_1)}{(v_2 - v_1)^3}\right).$$

– If $v_1 = v_3 \neq v_2 = v_4$, then

$$\frac{1}{|K|}\int_K \zeta(v_{\mathcal{T}}(x))\,\mathrm{d}x = 6\left(\frac{\mathcal{I}^2(\zeta)(v_1) + \mathcal{I}^2(\zeta)(v_2)}{(v_2 - v_1)^2} - 2\frac{\mathcal{I}^3(\zeta)(v_2) - \mathcal{I}^3(\zeta)(v_1)}{(v_2 - v_1)^3}\right).$$

– If $v_1 \neq v_2 \neq v_3 = v_4$, then

$$\frac{1}{|K|}\int_K \zeta(v_{\mathcal{T}}(x))\,\mathrm{d}x = 6\bigg(\frac{\mathcal{I}^2(\zeta)(v_3)}{(v_3-v_1)(v_3-v_2)} - \frac{\mathcal{I}^3(\zeta)(v_3)}{(v_3-v_1)(v_3-v_2)^2} - \frac{\mathcal{I}^3(\zeta)(v_3)}{(v_3-v_2)(v_3-v_1)^2} + \frac{\mathcal{I}^3(\zeta)(v_1)}{(v_1-v_3)^2(v_1-v_2)} + \frac{\mathcal{I}^3(\zeta)(v_2)}{(v_2-v_3)^2(v_2-v_1)}\bigg).$$

– In the case where all the values v_i are distinct, we get

$$\frac{1}{|K|}\int_K \zeta(v_{\mathcal{T}}(x))\,\mathrm{d}x = 6\sum_{i=1}^4 \frac{\mathcal{I}^3(\zeta)(v_i)}{\prod_{j\in\{1,2,3,4\}\setminus\{i\}}(v_i-v_j)}.$$

Since the numerical results are computed in the particular case where $\zeta = \psi'_\alpha$, we have $\mathcal{I}(\zeta) = \psi_\alpha$. Let us now provide expressions for $\mathcal{I}^2(\zeta) = \mathcal{I}(\psi_\alpha)$ and $\mathcal{I}^3(\zeta) = \mathcal{I}^2(\psi_\alpha)$.

We introduce the so-called "expintegral" function, defined by

$$\forall x \le 0,\ \mathrm{Ei}(x) := \int_{-\infty}^x \frac{e^t}{t}\,\mathrm{d}t \text{ and } \forall x > 0,\ \mathrm{Ei}(x) := \lim_{h\to 0, h>0}\bigg(\int_{-\infty}^{-h}\frac{e^t}{t}\,\mathrm{d}t + \int_h^x \frac{e^t}{t}\,\mathrm{d}t\bigg).$$

We then have

$$\forall s\in[0,1),\ \mathcal{I}(\psi_\alpha)(s) = (1-s)\bigg(1-\exp\bigg(\frac{\alpha s}{1-s}\bigg)\bigg) + \frac{\alpha}{e^\alpha}\bigg(\mathrm{Ei}\bigg(\frac{\alpha}{1-s}\bigg) - \mathrm{Ei}(\alpha)\bigg),$$

with $\mathcal{I}(\psi_\alpha)(-s) = \mathcal{I}(\psi_\alpha)(s)$ for all $s\in[0,1)$, and

$$\forall s\in[0,1),\ \mathcal{I}^2(\psi_\alpha)(s) = s - \frac{s^2}{2} + \frac{(1-s)(1-s-\alpha)}{2}\exp\bigg(\frac{\alpha s}{1-s}\bigg) + \alpha\frac{s-1+\frac{\alpha}{2}}{e^\alpha}\bigg(\mathrm{Ei}\bigg(\frac{\alpha}{1-s}\bigg) - \mathrm{Ei}(\alpha)\bigg) + \frac{\alpha-1}{2},$$

with $\mathcal{I}^2(\psi_\alpha)(-s) = -\mathcal{I}^2(\psi_\alpha)(s)$ for all $s\in[0,1)$.

A.2. Equivalence of norms

Partial proofs of the following lemma are done in the literature, for example in [6], and we only provide the sketch of the proof for the sake of completeness.

Lemma A.2. *For any $r\in(1,+\infty)$, there exists $C_7^{(r)} > 0$, also depending on d, such that, for any $v\in\mathbb{R}^{\mathcal{N}}$,*

$$\frac{1}{C_7^{(r)}}\|\Pi_{\mathcal{T}} v\|_r \le \|\Pi_{\mathcal{M}} v\|_r \le C_7^{(r)}\|\Pi_{\mathcal{T}} v\|_r, \tag{A.1}$$

and there exists $C_8^{(\theta_{\mathcal{T}},r)} > 0$, which is increasing with respect to $\theta_{\mathcal{T}}$ and also depending on d, such that, for any $v\in\mathbb{R}^{\mathcal{N}}$,

$$\frac{1}{C_8^{(\theta_{\mathcal{T}},r)}}\|\nabla\Pi_{\mathcal{T}} v\|_r \le \|v\|_{1,r,\mathcal{M}} \le C_8^{(\theta_{\mathcal{T}},r)}\|\nabla\Pi_{\mathcal{T}} v\|_r \tag{A.2}$$

where $\|\cdot\|_{1,r,\mathcal{M}}$ is defined by (3.16).

Proof. We denote by K_0 the reference simplex with vertices 0 and all the extremities of the canonical unit vectors. Let μ_K be the affine mapping which transforms K_0 into $K \in \mathcal{T}$ with vertices $(z_{i_0}, \dots, z_{i_d})$. Then $\Pi_{\mathcal{T}} v \circ \mu_K$ is an affine function on K_0 with values $(v_{i_0}, \dots, v_{i_d})$ at the vertices of K_0. By equivalence of norms in finite dimension, there exists $C > 0$ only depending on d and r such that

$$\frac{1}{C} \int_{K_0} |\Pi_{\mathcal{T}} v(\mu_K(x))|^r \, \mathrm{d}x \le \sum_{k=0}^{d} |v_{i_k}|^r \le C \int_{K_0} |\Pi_{\mathcal{T}} v(\mu_K(x))|^r \, \mathrm{d}x.$$

Then (A.1) results from

$$\int_K |\Pi_{\mathcal{T}} v(x)|^r \, \mathrm{d}x = \frac{|K|}{|K_0|} \int_{K_0} |\Pi_{\mathcal{T}} v(\mu_K(x))|^r \, \mathrm{d}x.$$

Denoting by $D\mu_K$ the Jacobian matrix of the change of variable, we have

$$\int_K |\nabla \Pi_{\mathcal{T}} v(x)|^r \, \mathrm{d}x = \frac{|K|}{|K_0|} \int_{K_0} |D\mu_K \nabla(\Pi_{\mathcal{T}} v \circ \mu_K)(x))|^r \, \mathrm{d}x,$$

where $\nabla(\Pi_{\mathcal{T}} v \circ \mu_K)(x)$ is the constant vector with components $(v_{i_k} - v_{i_0})_{k=1,\dots,d}$. Remarking that

$$|\nabla(\Pi_{\mathcal{T}} v \circ \mu_K)(x)| = \left(\sum_{k=1}^{d} (v_{i_k} - v_{i_0})^2 \right)^{1/2},$$

equation (A.2) follows from the equivalence of norms in finite dimension spaces, from a bound of the coefficients of $D\mu_K$ by h_K, from $h_K \le d_{ij} \theta_K$ and from a bound of the coefficients of $(D\mu_K)^{-1}$ by $h_K^{d-1}/|K|$ involving θ_K. □

The following lemma and its proof can be found in [6].

Lemma A.3. *For any $1 \le r \le \infty$, there exists $C_9(r)$ such that, for all $v = (v_i)_{i \in \mathcal{N}} \in \mathbb{R}^{\mathcal{N}}$,*

$$\|\Pi_{\mathcal{T}} v - \Pi_{\mathcal{M}} v\|_r \le C_9(r) h_{\mathcal{T}} \|\nabla \Pi_{\mathcal{T}} v\|_r.$$

Proof. Writing $|\Pi_{\mathcal{T}} v(x) - \Pi_{\mathcal{M}} v(x)|^r = |\nabla \Pi_{\mathcal{T}} v(x) \cdot (x - z_i)|^r$ for $x \in \omega_{i,K}$ and $i \in \mathcal{N}_K$, the conclusion follows from the upper bound of $|x - z_i|$ by $h_{\mathcal{T}}$ and the integration on $\omega_{i,K}$. □

A.3. Comparison lemmas

The next lemma plays an important role in the convergence properties of Scheme (3.8).

Lemma A.4. *For any $(a, b) \in (-1, 1)^2$ with $a \ne b$ and for any $q \in [0, 1)$, we have*

$$\min_{s \in I(a,b)} \psi'(s) \le \frac{\left(\widetilde{\psi}_q(a) - \widetilde{\psi}_q(b) \right)^2}{(a - b)(\psi_q(a) - \psi_q(b))} \le \max_{s \in I(a,b)} \psi'(s), \tag{A.3}$$

where the function ψ is defined by (2.2) and the functions ψ_q, $\widetilde{\psi}_q$ are defined by (3.9).

Proof. Note that (A.3) is proved for $q = 0$ in the proof of Lemma 3.10, since it holds $\widetilde{\psi}_0 = \beta$ and $\psi_0 = \mathrm{Id}$.

Let $(a, b) \in (-1, 1)^2$ and $q \in [0, 1)$. We assume that $a \le b$. Let us begin with the right inequality of (A.3).

$$\widetilde{\psi}_q(b) - \widetilde{\psi}_q(a) = \int_a^b \widetilde{\psi}_q'(t) \, \mathrm{d}t = \int_a^b \sqrt{\psi'(t)} \sqrt{\psi_q'(t)} \, \mathrm{d}t.$$

Consequently, we get, owing to the Cauchy–Schwarz inequality,

$$\left(\widetilde{\psi}_q(b)-\widetilde{\psi}_q(a)\right)^2 \le \int_a^b \psi'(t)\,\mathrm{d}t \int_a^b \psi_q'(t)\,\mathrm{d}t \le (\psi(b)-\psi(a))(\psi_q(b)-\psi_q(a)) \le \max_{s\in I(a,b)} \psi'(s)(b-a)(\psi_q(b)-\psi_q(a)),$$

which proves the right inequality of (A.3). Let us now turn to the left inequality of (A.3). We prove this inequality considering the different possible cases for $-1 \le a \le b \le 1$ (which are $q \le a \le b$, $-q \le a \le q \le b$, $a \le -q \le q \le b$, $-q \le a \le b \le q$, $a \le -q \le b \le q$ and $a \le b \le -q$). In the next computations, we denote, for short, by $\underline{\psi}'_{[a,b]} := \min_{s\in I(a,b)} \psi'(s)$.

Case 1. Let us assume that $q \le a \le b$, which also handles the case $a \le b \le -q$ by symmetry. We have

$$\begin{cases} \widetilde{\psi}_q(a) = \sqrt{\psi'(q)}(\beta(a)-\beta(q)) + \psi(q) \text{ and } \psi_q(a) = \psi(q) + \psi'(q)(a-q), \\ \widetilde{\psi}_q(b) = \sqrt{\psi'(q)}(\beta(b)-\beta(q)) + \psi(q) \text{ and } \psi_q(b) = \psi(q) + \psi'(q)(b-q). \end{cases}$$

We then obtain, writing $\beta(a)-\beta(b) = \sqrt{\psi'(c)}(a-b)$ with $c \in I(a,b)$,

$$\left(\widetilde{\psi}_q(a)-\widetilde{\psi}_q(b)\right)^2 = \psi'(q)(\beta(a)-\beta(b))^2 = \psi'(q)\psi'(c)(a-b)^2.$$

Since it holds $\psi_q(a)-\psi_q(b) = \psi'(q)(a-b)$, we obtain

$$\left(\widetilde{\psi}_q(a)-\widetilde{\psi}_q(b)\right)^2 = \psi'(c)(\psi_q(a)-\psi_q(b))(a-b) \ge \underline{\psi}'_{[a,b]}(\psi_q(a)-\psi_q(b))(a-b).$$

Case 2. Let us assume $-q \le a \le q \le b$, which also handles the case $a \le -q \le b \le q$ by symmetry. In this case we have

$$\begin{cases} \widetilde{\psi}_q(a) = \psi(a) & \text{and } \psi_q(a) = \psi(a), \\ \widetilde{\psi}_q(b) = \psi(q) + \sqrt{\psi'(q)}(\beta(b)-\beta(q)) & \text{and } \psi_q(b) = \psi(q) + \psi'(q)(b-q). \end{cases}$$

We have $\left(\widetilde{\psi}_q(a)-\widetilde{\psi}_q(b)\right)^2 - \underline{\psi}'_{[a,b]}(b-a)(\psi_q(b)-\psi_q(a)) = T_1+T_2+T_3+T_4$ with

$$\begin{aligned} T_1 &= \psi'(q)(\beta(b)-\beta(a))^2 - \underline{\psi}'_{[a,b]}\psi'(q)(b-q)^2, \\ T_2 &= \sqrt{\psi'(q)}(\beta(b)-\beta(q))(\psi(q)-\psi(a)) - \underline{\psi}'_{[a,b]}(b-q)(\psi(q)-\psi(a)), \\ T_3 &= \sqrt{\psi'(q)}(\beta(b)-\beta(q))(\psi(q)-\psi(a)) - \underline{\psi}'_{[a,b]}\psi'(q)(b-q)(q-a), \\ T_4 &= (\psi(q)-\psi(a))^2 - \underline{\psi}'_{[a,b]}(q-a)(\psi(q)-\psi(a)). \end{aligned}$$

Using the fact that $q \in I(a,b)$ and writing $\beta(b)-\beta(q) = \sqrt{\psi'(c)}(b-q)$ with $c \in I(q,b) \subset I(a,b)$, we have

$$\psi'(q)(\beta(b)-\beta(q))^2 = \psi'(q)\psi'(c)(b-q)^2 \ge \underline{\psi}'_{[a,b]}\psi'(q)(b-q)^2,$$

which proves that $T_1 \ge 0$. We can also write

$$\sqrt{\psi'(q)}(\beta(b)-\beta(q))(\psi(q)-\psi(a)) = \sqrt{\psi'(q)}\sqrt{\psi'(c)}(b-q)(\psi(q)-\psi(a)) \ge \underline{\psi}'_{[a,b]}(b-q)(\psi(q)-\psi(a)),$$

hence proving that $T_2 \ge 0$. Owing to $\psi(q)-\psi(a) = \psi'(e)(q-a)$ with $e \in I(a,q) \subset I(a,b)$, and to $\psi'(c) \ge \psi'(q)$ since $0 \le q \le c \le b$, we have

$$\sqrt{\psi'(q)}(\beta(b)-\beta(q))(\psi(q)-\psi(a)) = \sqrt{\psi'(q)}\sqrt{\psi'(c)}(b-q)\psi'(e)(q-a)$$
$$\geq \psi'(q)(b-q)\psi'(e)(q-a) \geq \underline{\psi}'_{[a,b]}\psi'(q)(b-q)(q-a),$$

which shows that $T_3 \geq 0$. We now write

$$(\psi(q)-\psi(a))^2 = \psi'(e)(q-a)(\psi(q)-\psi(a)) \geq \underline{\psi}'_{[a,b]}(q-a)(\psi(q)-\psi(a)),$$

hence concluding that $T_4 \geq 0$, which completes the study of Case 2.

Case 3. $a \leq -q \leq q \leq b$. In this case, we have $\underline{\psi}'_{[a,b]} = 1$ and

$$\begin{cases} \widetilde{\psi}_q(a) = -\psi(q) + \sqrt{\psi'(q)}(\beta(a)+\beta(q)) \text{ and } \psi_q(a) = -\psi(q) + \psi'(q)(a+q), \\ \widetilde{\psi}_q(b) = \psi(q) + \sqrt{\psi'(q)}(\beta(b)-\beta(q)) \quad \text{and } \psi_q(b) = \psi(q) + \psi'(q)(b-q). \end{cases}$$

Since

$$\begin{aligned} \widetilde{\psi}_q(b) - \widetilde{\psi}_q(a) &= 2\psi(q) + \sqrt{\psi'(q)}(\beta(b)-\beta(q)-(\beta(a)+\beta(q))) \\ b-a &= 2q + b - q - (a+q), \\ \psi_q(b) - \psi_q(a) &= 2\psi(q) + \psi'(q)(b-q-(a+q)), \end{aligned}$$

we compute $\left(\widetilde{\psi}_q(b) - \widetilde{\psi}_q(a)\right)^2 - (b-a)(\psi_q(b)-\psi_q(a)) = 4T_1 + 2T_2 + 2T_3 + T_4$ with

$$\begin{aligned} T_1 &= \psi(q)^2 - q\psi(q), \\ T_2 &= \psi(q)\sqrt{\psi'(q)}(\beta(b)-\beta(q)-(\beta(a)+\beta(q))) - q\psi'(q)(b-q-(a+q)), \\ T_3 &= \psi(q)\sqrt{\psi'(q)}(\beta(b)-\beta(q)-(\beta(a)+\beta(q))) - \psi(q)(b-q-(a+q)), \\ T_4 &= \psi'(q)(\beta(b)-\beta(q)-(\beta(a)+\beta(q)))^2 - \psi'(q)(b-q-(a+q))^2. \end{aligned}$$

The property $\psi(q) \geq q$ immediately implies that $T_1 \geq 0$.
Owing to $\beta(b)-\beta(q) \geq \sqrt{\psi'(q)}(b-q)$ and $-(\beta(a)+\beta(q)) \geq \sqrt{\psi'(q)}(-(a+q))$ which gives

$$\beta(b)-\beta(q)-(\beta(a)+\beta(q)) \geq \sqrt{\psi'(q)}(b-q-(a+q)) \geq 0.$$

Multiplying by $\psi(q)\sqrt{\psi'(q)} \geq q\sqrt{\psi'(q)} \geq 0$, we obtain $T_2 \geq 0$. Since $\sqrt{\psi'(q)} \geq 1$, we have

$$\sqrt{\psi'(q)}(\beta(b)-\beta(q)-(\beta(a)+\beta(q))) \geq b-q-(a+q)) \geq 0,$$

which gives $T_3 \geq 0$, and the inequality

$$\beta(b)-\beta(q)-(\beta(a)+\beta(q)) \geq b-q-(a+q)) \geq 0,$$

implies that $T_4 \geq 0$, hence concluding the study of Case 3.

Case 4. Let us assume that $-q \leq a \leq b \leq q$. We have

$$\begin{cases} \widetilde{\psi}_q(a) = \psi_q(a) = \psi(a) \\ \widetilde{\psi}_q(b) = \psi_q(b) = \psi(b). \end{cases}$$

We then obtain

$$\left(\widetilde{\psi}_q(a) - \widetilde{\psi}_q(b)\right)^2 = (\psi(a)-\psi(b))^2 = (\psi(a)-\psi(b))(\psi(a)-\psi(b)).$$

Now we write $(\psi(a)-\psi(b)) = \psi'(c)(a-b)$ with $c \in I(a,b)$. Using the monotonicity of ψ we have

$$(\psi(a)-\psi(b))(\psi(a)-\psi(b)) = \psi'(c)(a-b)(\psi(a)-\psi(b)) \geq \underline{\psi}'_{[a,b]}(a-b)(\psi(a)-\psi(b)),$$

which concludes the proof of the lemma.

□

The following lemma is needed for the proof of the convergence to a weak solution.

Lemma A.5. *For any $\varepsilon \in (0, \frac{7}{4})$, there exists $\nu_\varepsilon > 0$ only depending on ε such that, for any $a, b \in (-1, 1)$, it holds*

$$|b-a|(\psi'(a)+\psi'(b)) \leq \nu_\varepsilon |\beta(b)-\beta(a)| \left(2+|\beta(a)|^{1+\varepsilon}+|\beta(b)|^{1+\varepsilon}\right), \tag{A.4}$$

where ψ is defined by (2.2) and β is defined by (2.10).

Proof. We first remark that, if $a \leq 0 \leq b$, setting $c = \max(|a|, |b|) \in [0, 1)$, then

$$|b-a|(\psi'(a)+\psi'(b)) \leq 4c\psi'(c) \leq 4|c-0|(\psi'(c)+\psi'(0)),$$

and

$$|\beta(c)-\beta(0)| \left(2+|\beta(c)|^{1+\varepsilon}+|\beta(0)|^{1+\varepsilon}\right) \leq |\beta(b)-\beta(a)| \left(2+|\beta(a)|^{1+\varepsilon}+|\beta(b)|^{1+\varepsilon}\right).$$

It is therefore sufficient to prove (A.4) in the case $0 \leq a \leq b$, which includes the case where $0 \leq a := 0 \leq b := c$.

Denoting by $A = a/(1-a)$ and $B = b/(1-b)$, let us prove that, for any $\nu \in (0, \frac{1}{4})$, we have

$$\left(e^{\nu B}-e^{\nu A}\right)\left(e^{(1+\nu)B}+e^{(1+\nu)A}\right) \leq \left(e^{\left(\frac{1}{2}-\nu\right)B}-e^{\left(\frac{1}{2}-\nu\right)A}\right)\left(e^{\left(\frac{1}{2}+3\nu\right)B}+e^{\left(\frac{1}{2}+3\nu\right)A}\right). \tag{A.5}$$

Indeed, we have

$$\left(e^{\left(\frac{1}{2}-\nu\right)B}-e^{\left(\frac{1}{2}-\nu\right)A}\right)\left(e^{\left(\frac{1}{2}+3\nu\right)B}+e^{\left(\frac{1}{2}+3\nu\right)A}\right)-\left(e^{\nu B}-e^{\nu A}\right)\left(e^{(1+\nu)B}+e^{(1+\nu)A}\right) = e^{(1+2\nu)A} f_\nu(B-A),$$

with, for all $x \in [0, +\infty)$,

$$f_\nu(x) := \left(e^{\left(\frac{1}{2}-\nu\right)x}-1\right)\left(e^{\left(\frac{1}{2}+3\nu\right)x}+1\right)-(e^{\nu x}-1)\left(e^{(1+\nu)x}+1\right).$$

After simplification, we obtain

$$f_\nu(x) = e^{(1+\nu)x}-e^{\nu x}-e^{\left(\frac{1}{2}+3\nu\right)x}+e^{\left(\frac{1}{2}-\nu\right)x} = e^{(1+\nu)x}(1-e^{-x})-e^{\left(\frac{1}{2}+3\nu\right)x}(1-e^{-4\nu x}).$$

Using the fact that $\nu \in (0, \frac{1}{4})$ implies $1+\nu \geq \frac{1}{2}+3\nu$, we can write

$$e^{(1+\nu)x} \geq e^{\left(\frac{1}{2}+3\nu\right)x} \text{ and } 1-e^{-x} \geq 1-e^{-4\nu x}.$$

Therefore we get that $f_\nu(x) \geq 0$, which proves (A.5). Following the computations of Lemma 2.6, we recall that there exists $\mu > 0$ only depending on $\nu > 0$ such that we have

$$\begin{aligned} \psi'(a) \leq \mu e^{(1+\nu)A} \text{ and } \psi'(b) \leq \mu e^{(1+\nu)B}, \\ b-a \leq \mu\left(e^{\nu B}-e^{\nu A}\right), \\ e^{\left(\frac{1}{2}-\nu\right)B}-e^{\left(\frac{1}{2}-\nu\right)A} \leq \mu(\beta(b)-\beta(a)), \end{aligned}$$

and

$$e^{\left(\frac{1}{2}+3\nu\right)A} \leq \mu|\beta(a)|^{1+7\nu}+1 \text{ and } e^{\left(\frac{1}{2}+3\nu\right)B} \leq \mu|\beta(b)|^{1+7\nu}+1.$$

Gathering the previous inequalities and (A.5) provides (A.4), letting $\varepsilon = 7\nu$ with $\nu \in \left(0, \frac{1}{4}\right)$. □

Lemma A.6. *Let $\overline{u} \in \mathcal{S}_d(\Omega)$ be an entropy solution of problems (1.1) and (1.2) in the sense of Definition 3.1. Then, for for any $k > 0$ and for any $\phi \in C_c^\infty(\Omega)$, we have*

$$\int_{|\overline{u}-\phi|\le k} \Lambda\nabla\overline{u}\cdot\nabla(\overline{u}-\phi)\,\mathrm{d}x = \int_\Omega T_k(\overline{u}-\phi)f\,\mathrm{d}x. \tag{A.6}$$

As a consequence, we obtain

$$\int_\Omega \Lambda\nabla\overline{u}\cdot\nabla T(\overline{u}-\phi)\,\mathrm{d}x = \int_\Omega T(\overline{u}-\phi)f\,\mathrm{d}x, \tag{A.7}$$

for any $\phi \in C_c^\infty(\Omega)$ and for any $T \in \mathcal{F}$, where $\mathcal{F}$ is defined in Remark 3.2.

Proof. We first use Lemma 3.3 of [3], which proves by a density argument that (3.1) also holds for any $\phi \in H_0^1(\Omega)\cap L^\infty(\Omega)$. We therefore let $\phi = 2T_h(\overline{u}) - \widetilde{\phi}$ in (3.1), for given $h > 0$ and $\widetilde{\phi} \in C_c^\infty(\Omega)$. This gives

$$\int_\Omega \Lambda\nabla\overline{u}\cdot\nabla T_k\Big(\overline{u}-2T_h(\overline{u})+\widetilde{\phi}\Big)\,\mathrm{d}x \le \int_\Omega T_k\Big(\overline{u}-2T_h(\overline{u})+\widetilde{\phi}\Big)f\,\mathrm{d}x. \tag{A.8}$$

Let $M = k + \|\widetilde{\phi}\|_\infty$. For $h > M$, we obtain that:

- $T_k(\overline{u}-2T_h(\overline{u})+\widetilde{\phi}) = \overline{u}+2h+\widetilde{\phi}$ for $\left|\overline{u}+2h+\widetilde{\phi}\right| \le k$,
- $T_k(\overline{u}-2T_h(\overline{u})+\widetilde{\phi}) = -\overline{u}+\widetilde{\phi}$ for $\left|-\overline{u}+\widetilde{\phi}\right| \le k$,
- $T_k(\overline{u}-2T_h(\overline{u})+\widetilde{\phi}) = \overline{u}-2h+\widetilde{\phi}$ for $\left|\overline{u}-2h+\widetilde{\phi}\right| \le k$,
- otherwise $T_k\Big(\overline{u}-2T_h(\overline{u})+\widetilde{\phi}\Big) = \pm k$,

and that

$$T_k\Big(\overline{u}-2T_h(\overline{u})+\widetilde{\phi}\Big) = T_k\Big(-\overline{u}+\widetilde{\phi}\Big) \text{ if } |\overline{u}| \le 2h-M. \tag{A.9}$$

This leads to

$$\nabla T_k\Big(\overline{u}-2T_h(\overline{u})+\widetilde{\phi}\Big) = \nabla T_k\Big(-\overline{u}+\widetilde{\phi}\Big) + \nabla T_k\Big(\overline{u}+2h+\widetilde{\phi}\Big) + \nabla T_k\Big(\overline{u}-2h+\widetilde{\phi}\Big).$$

We thus obtain

$$\int_\Omega \Lambda\nabla\overline{u}\cdot\nabla T_k\Big(\overline{u}-2T_h(\overline{u})+\widetilde{\phi}\Big)\,\mathrm{d}x = \int_\Omega \Lambda\nabla\overline{u}\cdot\nabla T_k\Big(-\overline{u}+\widetilde{\phi}\Big)\,\mathrm{d}x + R_h,$$

with

$$|R_h| \le \overline{\lambda}\int_{2h-M<|\overline{u}|<2h+M} |\nabla\overline{u}|\Big(|\nabla\overline{u}| + \left|\nabla\widetilde{\phi}\right|\Big)\,\mathrm{d}x.$$

Applying Lemma A.7, we get that

$$\lim_{h\to\infty} R_h = 0.$$

Besides, we get from (A.9) that

$$\left|\int_\Omega T_k\Big(\overline{u}-2T_h(\overline{u})+\widetilde{\phi}\Big)f\,\mathrm{d}x - \int_\Omega T_k\Big(-\overline{u}+\widetilde{\phi}\Big)f\,\mathrm{d}x\right| \le 2k\int_{|\overline{u}|\ge 2h-M} |f|\,\mathrm{d}x.$$

By dominated convergence, we get that

$$\lim_{h\to\infty}\int_{|\overline{u}|\ge 2h-M} |f|\,\mathrm{d}x = 0.$$

Letting $h \to \infty$ in (A.8), we therefore obtain

$$\int_\Omega \Lambda \nabla \overline{u} \cdot \nabla T_k\Big(-\overline{u} + \widetilde{\phi}\Big)\,\mathrm{d}x \le \int_\Omega T_k\Big(-\overline{u} + \widetilde{\phi}\Big) f\,\mathrm{d}x,$$

which, in addition to (3.1) with $\phi = \widetilde{\phi}$, provides (A.6).

We then deduce (A.7), using for any $T \in \mathcal{F}$ the relation

$$\forall s > 0,\ T(s) = \int_0^{+\infty} (-T''(k)) T_k(s)\,\mathrm{d}k,$$

multiplying (A.6) by $(-T''(k))$ and integrating with respect to k (as it is suggested in [3]). □

Lemma A.7. *Let $\overline{u} \in \mathcal{S}_d(\Omega)$ be an entropy solution of problems (1.1) and (1.2) in the sense of Definition 3.1. Then, for all $k > 0$, we have*

$$\lim_{h \to +\infty} \int_{h-k < |\overline{u}| \le h+k} |\nabla \overline{u}|^2\,\mathrm{d}x = 0. \tag{A.10}$$

Proof. As in the proof of Lemma A.6, we use the fact that (3.1) also holds for any $\phi \in H_0^1(\Omega) \cap L^\infty(\Omega)$. Letting, for given $k, h > 0$, $\phi = T_h(\overline{u})$ in (3.1), we get

$$\int_\Omega \Lambda \nabla \overline{u} \cdot \nabla T_k(\overline{u} - T_h(\overline{u}))\,\mathrm{d}x \le \int_\Omega f T_k(\overline{u} - T_h(\overline{u}))\,\mathrm{d}x.$$

Using $\nabla \overline{u} = \nabla T_k(\overline{u} - T_h(\overline{u}))$ for *a.e.* x such that $\nabla T_k(\overline{u} - T_h(\overline{u}))(x) \neq 0$, we get, denoting by $E_h = \{x \in \Omega, h < |\overline{u}(x)| \le h + k\}$,

$$\underline{\lambda} \|\nabla \overline{u}\|^2_{L^2(E_h)} \le \int_\Omega f T_k(\overline{u} - T_h(\overline{u}))\,\mathrm{d}x,$$

which gives

$$\underline{\lambda} \|\nabla \overline{u}\|^2_{L^2(E_h)} \le \int_{|\overline{u}| > h} k|f|\,\mathrm{d}x.$$

By dominated convergence, we get

$$\lim_{h \to +\infty} \int_{|\overline{u}| > h} k|f|\,\mathrm{d}x = 0,$$

and therefore we obtain

$$\lim_{h \to +\infty} \underline{\lambda} \|\nabla \overline{u}\|^2_{L^2(E_h)} = 0. \tag{A.11}$$

Note that (A.11) implies, replacing h by $h - k$, that

$$\lim_{h \to +\infty} \int_{h < |\overline{u}| \le h+k} |\nabla \overline{u}|^2\,\mathrm{d}x = \lim_{h \to +\infty} \int_{h-k < |\overline{u}| \le h} |\nabla \overline{u}|^2\,\mathrm{d}x = 0,$$

hence providing (A.10). □

Acknowledgements. The authors thank Thierry Gallouët and Alain Prignet for inspiring discussions.

References

[1] R.A. Adams, Sobolev Spaces. *Pure and Applied Mathematics*. Vol. 65. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London (1975).

[2] A. Ait Hammou Oulhaj, C. Cancès and C. Chainais-Hillairet, Numerical analysis of a nonlinearly stable and positive control volume finite element scheme for Richards equation with anisotropy. *ESAIM: M2AN* **52** (2018) 1533–1567.

[3] P. Bénilan, L. Boccardo, T. Gallouët, R. Gariepy, M. Pierre and J.L. Vázquez, An L^1-theory of existence and uniqueness of solutions of nonlinear elliptic equations. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* **22** (1995) 241–273.
[4] L. Boccardo and T. Gallouët, Nonlinear elliptic and parabolic equations involving measure data. *J. Funct. Anal.* **87** (1989) 149–169.
[5] L. Boccardo, T. Gallouët and J.L. Vázquez, Nonlinear elliptic equations in $\mathbf{R}^N$ without growth restrictions on the data. *J. Differ. Equ.* **105** (1993) 334–363.
[6] C. Cancès and C. Guichard, Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations. *Math. Comp.* **85** (2016) 549–580.
[7] C. Cancès, M. Ibrahim and M. Saad, Positive nonlinear CVFE scheme for degenerate anisotropic Keller-Segel system. *SMAI J. Comput. Math.* **3** (2017) 1–28.
[8] J. Casado-Díaz, T. Chacón Rebollo, V. Girault, M. Gómez Mármol and F. Murat, Finite elements approximation of second order linear elliptic equations in divergence form with right-hand side in L^1. *Numer. Math.* **105** (2007) 337–374.
[9] P.G. Ciarlet, The finite element method for elliptic problems. In: Studies in Mathematics and its Applications. Vol. 4. North-Holland Publishing Co., Amsterdam-New York-Oxford (1978) xix+530.
[10] G. Dal Maso, F. Murat, L. Orsina and A. Prignet, Renormalized solutions of elliptic equations with general measure data. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* **28** (1999) 741–808.
[11] A. Dall'Aglio, A remark on the entropy solutions. Personal communication (1996).
[12] K. Deimling, Nonlinear Functional Analysis. Springer-Verlag, Berlin (1985).
[13] J. Droniou, Solving convection-diffusion equations with mixed, neumann and fourier boundary conditions and measures as data, by a duality method. *Adv. Differ. Equ.* **5** (2000) 1341–1396.
[14] J. Droniou, Non-coercive linear elliptic problems. *Potential Anal.* **17** (2002) 181–203.
[15] J. Droniou, T. Gallouët and R. Herbin, A finite volume scheme for a noncoercive elliptic equation with measure data. *SIAM J. Numer. Anal.* **41** (2003) 1997–2031.
[16] A. Ern and J.-L. Guermond, Theory and Practice of Finite Elements. Vol. 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York (2004).
[17] R. Eymard, T. Gallouët and R. Herbin, Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI: a scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.* **30** (2010) 1009–1043.
[18] R. Eymard, T. Gallouët, C. Guichard, R. Herbin and R. Masson, TP or not TP, that is the question. *Comput. Geosci.* **18** (2014) 285–296.
[19] P. Fabrie and T. Gallouët, Modelling wells in porous media flow. *Math. Models Methods Appl. Sci.* **10** (2000) 673–709.
[20] T. Gallouët and R. Herbin, Existence of a solution to a coupled elliptic system. *Appl. Math. Lett.* **7** (1994) 49–55.
[21] T. Gallouët and A. Monier, On the regularity of solutions to elliptic equations. *Rend. Mat. Appl.* **19** (1999) 471–488.
[22] N.G. Meyers, An L^p-estimate for the gradient of solutions of second order elliptic divergence equations. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* **17** (1963) 189–206.
[23] A. Prignet, Remarks on existence and uniqueness of solutions of elliptic problems with right-hand side measures. *Rend. Mat. Appl.* **15** (1995) 321–337.
[24] J. Serrin, Pathological solutions of elliptic differential equations. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* **18** (1964) 385–387.
[25] G. Stampacchia, Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus. *Ann. Inst. Fourier (Grenoble)* **15** (1965) 189–258.