

**AN IMPLICIT SCHEME TO SOLVE A SYSTEM OF ODES
ARISING FROM THE SPACE DISCRETIZATION OF NONLINEAR DIFFUSION
EQUATIONS***

ÉRIC BOILLAT¹

Abstract. In this article, we consider the initial value problem which is obtained after a space discretization (with space step h) of the equations governing the solidification process of a multicomponent alloy. We propose a numerical scheme to solve numerically this initial value problem. We prove an error estimate which is not affected by the step size h chosen in the space discretization. Consequently, our scheme provides global convergence without any stability condition between h and the time step size τ . Moreover, it is not of excessive algorithmic complexity since it does not require more than one resolution of a linear system at each time step.

Mathematics Subject Classification. 65L05, 65L80, 65N30.

Received: April 10, 2000. Revised: February 12, 2001.

1. INTRODUCTION

Consider an isotropic material composed by $m \geq 1$ chemical species and contained in a polyhedral domain $\Omega \subset \mathbb{R}^3$. During the time interval $[0, T]$, the thermodynamical state of this system is described by two \mathbb{R}^m -valued mapping w and ψ defined on $Q_T = (0, T) \times \Omega$. The m components of w are the *conserved variables* and the components of ψ are called *generalized potentials*. Their physical interpretation is as follows: $w_1(t, x)$ is the specific enthalpy at time $t \in (0, T)$ and at point $x \in \Omega$, $\psi_1(t, x)^{-1}$ is the temperature, $w_j(t, x)$ is the concentration of the j th chemical specie and the product $\psi_{j+1}(t, x)\psi_1(t, x)^{-1}$ represents its chemical potential. The relation between the conserved variables and the generalized potentials is algebraic. It reads

$$\psi_j(t, x) = \frac{\partial \sigma}{\partial w_j}(w(t, x)), \quad (t, x) \in \overline{Q_T}, \quad j = 1 \dots m, \quad (1.1)$$

where $\sigma : \mathbb{R}^m \rightarrow \mathbb{R}$ is a differentiable and concave function. The quantity $\sigma(w(t, x))$ is interpreted as the specific entropy at time t and at position x .

Keywords and phrases. Nonlinear diffusion equations, nonlinear parabolic problem, Chernoff scheme, implicit scheme for ODE's.

* This work is supported by the Swiss National Funds for Scientific Research.

¹ Department of Mathematics, EPFL, 1015 Lausanne, Switzerland. e-mail: eric.boillat@epfl.ch

If there are no convective motions in the domain Ω , it follows from a first order approximation of the theory of irreversible processes (see [4] or [7]) that the conserved variables satisfy the following evolution equation,

$$\partial_t w_j + \operatorname{div} \sum_{i=1}^{i=m} L_{ji} \nabla \psi_i = G_j(\psi), \text{ in } Q_T, j = 1 \dots m. \tag{1.2}$$

The quantities L_{ij} are diffusion coefficients. They are known functions of time, position and generalized potentials. According to the Onsager reciprocity principle [7], the $m \times m$ matrix (L_{ij}) is symmetric and positive definite. Let us stress that the L_{ij} 's are scalar quantities because of the isotropy hypothesis. For an anisotropic material, they would be 3×3 matrices. In relation (1.2), $G_j(\psi)$ is a source term due to chemical reactions. It depends on the generalized potentials. It is assumed that $G : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is a Lipschitz continuous mapping and the second principle of thermodynamics implies that

$$\psi \cdot G(\psi) \geq 0, \psi \in \mathbb{R}^m. \tag{1.3}$$

The equations (1.1–1.2) need to be completed by an initial condition for the conserved variables:

$$w(0, x) = w^0(x), \forall x \in \Omega \tag{1.4}$$

and by convenient boundary conditions for the generalized potentials. In the simplest case, homogeneous boundary conditions are imposed. In that case, the solution ψ to (1.1–1.2) has to be searched in $L^2(0, T; H_0^1(\Omega)^m)$ while w has to be of class $L^2(0, T; L^2(\Omega)^m)$ with a time derivative $\partial_t w$ in $L^2(0, T; H^{-1}(\Omega)^m)$. We refer to the article by Donnelly [8] where a proof can be found for the existence and uniqueness to the solution of (1.1–1.4) when the diffusion matrix (L_{ij}) is the identity and when the range of the mapping $w \in \mathbb{R}^m \mapsto \nabla \sigma(w)$ is the whole \mathbb{R}^m . An other proof of existence and uniqueness for the solution to (1.1–1.4) with more general diffusion matrices is also given in an article by Alt and Luckhaus [6]. Observe however that very few results are available when the gradient of σ fails to be surjective from \mathbb{R}^m into \mathbb{R}^m .

In this work, we will not consider the analysis of the continuous equations (1.1–1.4) any further. Our purpose is rather to concentrate on finite element space discretizations to (1.1–1.4) and, in particular, on the numerical integration of the obtained system of o.d.e.

For the sake of simplicity, we will expose the subject of this article by assuming that the diffusion coefficients matrix (L_{ij}) is the identity and that homogeneous Dirichlet boundary conditions are imposed on the potentials. A standard f.e.m. applied to (1.1–1.4) can be described as follows. Let $\{\mathcal{M}_h\}$ be a regular family of meshes of Ω made of tetrahedrons and satisfying an inverse condition [19]. The subscript $h > 0$ denotes the grid size of the mesh \mathcal{M}_h and we call $P_1 \dots P_N$ its interior nodes. To each interior node P_j , we associate the Delaunay cell $\mathcal{O}_j = \{P \in \Omega, |\operatorname{dist}(P, P_j) < \operatorname{dist}(P, P_k), k \neq j\}$. The Delaunay cells build a new mesh of Ω called dual mesh and denoted by \mathcal{M}'_h . The meshes \mathcal{M}_h and \mathcal{M}'_h are used to construct two finite dimensional spaces V_h and W_h . The first one contains the functions in $H_0^1(\Omega)^m$ that are piecewise linear on \mathcal{M}_h while the second one is made of all the functions in $L^2(\Omega)^m$ that are piecewise constant on \mathcal{M}'_h . For a given value of h , the finite element approximation for the solution (w, ψ) to (1.1–1.4) is defined as the pair $(u, \phi) \in C^1([0, T]; W_h) \times C^0([0, T]; V_h)$ solving equation (1.2) when it is tested against any element of V_h and equations (1.1), (1.4) when they are tested against any element of W_h . Using Green's theorem and numerical integration, we get the following system of algebraic and differential equations for u and ϕ . It reads

$$(\dot{u}(t), \xi)_{0,h} - (\phi(t), \xi)_1 = (g(u(t)), \xi)_{0,h}, \forall \xi \in V_h, \forall t \in (0, T), \tag{1.5}$$

$$(v, \phi(t))_{0,h} - (\partial S(u(t)), v)_0 = 0, \forall v \in W_h, \forall t \in [0, T], \tag{1.6}$$

$$u(0, \cdot) = u^0, \tag{1.7}$$

where the initial condition u^0 is the $L^2(\Omega)^m$ -projection of w^0 onto W_h . In equations (1.5–1.7), the notations $(\cdot, \cdot)_0$ and $(\cdot, \cdot)_1$ are used for the scalar products in $L^2(\Omega)^m$ and $H_0^1(\Omega)^m$ and $(\cdot, \cdot)_{0,h}$ is a numerical formula

for $(\cdot, \cdot)_0$. We define it as

$$(v, \xi)_{0,h} \equiv \sum_{j=1}^{j=N} |\mathcal{O}_j| v(P_j)^T \xi(P_j), \tag{1.8}$$

where $|\mathcal{O}_j|$ is the measure of the Delaunay cells \mathcal{O}_j associated to P_j . In equations (1.5) and (1.6), ∂S denotes the first variation of a functional S and g is a Niemicki operator. For any $v \in L^2(\Omega)^m$, $S(v) \in \mathbb{R}$ and $g(v) \in L^2(\Omega)^m$ are defined by

$$S(v) = \int_{\Omega} \sigma(v(x)) dx \quad \text{and} \quad g(v)(x) = G(\nabla \sigma(v(x))), \quad x \in \Omega. \tag{1.9}$$

Our goal is now to design proper algorithms for integrating the system of o.d.e. (1.5–1.7). We are thus not interested in establishing the convergence properties for the pair (u, ϕ) to the solution (w, ψ) to (1.1), (1.2) and (1.4) when h goes to zero. This question is addressed for linear cases in the book by Thomée [21]. We can also refer the reader to an article by Jerome and Rose [10], where a spatial Galerkin method for the diffusion equation governing the solidification of a pure material is analyzed.

The simplest way to integrate the system of o.d.e. (1.5–1.7) is the Forward-Euler method. Let $\tau > 0$ be a time step and let $u^n \in W_h$ be an approximation of the extensive variables at time $t_n = n\tau$. We compute the approximation $\phi^n \in V_h$ of the generalized potentials thanks to equation (1.6),

$$(v, \phi^n)_{0,h} = (\partial S(u^n), v)_0, \quad v \in W_h.$$

We then get an approximation $u^{n+1} \in W_h$ for $u(t_{n+1})$ by using the evolution equation (1.5),

$$(u^{n+1}, \xi)_0 = (u^n, \xi)_0 + \tau(\phi^n, \xi)_1 + \tau(g(u^n), \xi)_{0,h}, \quad \xi \in V_h$$

and we start again in the same way, computing ϕ^{n+1} and u^{n+2} . Unfortunately, solving problem (1.1–1.4) by combining the Forward-Euler method to the space discretization (1.5–1.7) has a major disadvantage. Even if the spatial discretization (1.5–1.7) is a converging method, letting the mesh size h and the time step τ tend to zero do not always provide convergence for the fully discrete solution (u^n, ϕ^n) . The reason is that the convergence properties of the Forward-Euler method are affected by h . To insure convergence of (u^n, ϕ^n) we actually have to respect a *stability condition*,

$$\rho(h)\tau \leq 1, \tag{1.10}$$

where $\rho(h)$ denotes a quantity proportional to the largest generalized eigenvalue of the rigidity matrix involved in (1.5). In usual situations, $\rho(h)$ grows like $\frac{1}{h^2}$ when h vanishes and condition (1.10) is therefore very restrictive. It compels τ to be so small that the number of operations to perform the numerical integration of (1.5–1.7) over the time interval $[0, T]$ is quite prohibitive for small h . By comparison, combining the finite element technique (1.5–1.7) with an integration algorithm whose convergence properties are *uniformly valid* in h would provide a global method that converges to the solution of (1.1–1.4) as soon as the discretization parameters h and τ tend to zero in arbitrary ways.

An other possibility to integrate (1.5–1.7) is the Backward-Euler method. This algorithm has the classical drawback to require the resolution of large system of *non-linear* equations at each time step. It has however been analyzed by Ciavaldini, Meyer, Jerome, Rose and Elliott in a particular case of the system (1.1–1.4) called *Stefan problem*. This problem describes the solidification processes of a pure material [1] and can be obtained by setting $m = 1$ and by choosing a piecewise quadratic entropy in (1.1–1.4). In that context, it has been proved in [9,10,15], or [3] that the Backward-Euler method approaches the enthalpy independently of h at order $\tau^{1/2}$ in the norm $L^\infty(0, T; H^{-1}(\Omega))$. A different approach has been used in [20]. The Backward-Euler method applied

to the Stefan problem is analyzed by mean of the semi-group theory and it is proved that the approximations for the temperature converge at order τ in $L^2(0, T; L^2(\Omega))$.

Recently, an other idea has been developed by Berger, Brezis, Rogers, Magenes, Nocketto, Verdi, Paolini and Sacchi see [2, 12, 13] and [18]. They integrate numerically the Stefan problem with an algorithm based on the Chernoff non-linear formula. Unlike the Backward-Euler method, it only requires the resolution of a *linear* system at each time step. Moreover, the aforementioned authors show that it also converges independently of h . They unfortunately get a suboptimal uniform order of $\tau^{1/4}$. The Chernoff algorithm has also been generalized to perform efficiently in context that are different from the Stefan problem. Let us quote a series of paper by Jäger and Kačur (see [22] and the references therein) where a variant of the Chernoff formula (called relaxation scheme) is applied to the porous medium problem describing the evolution of the density of a liquid flowing in a soil.

In the present article, our idea is to generalize the Chernoff algorithm and to apply it to the system of o.d.e. (1.5–1.7). Our main output will be that the result obtained by Magenes *et al.* is still valid in this more general context. We will show that the Chernoff algorithm integrates (1.5–1.7) at order $\tau^{1/4}$ uniformly in h . This result will actually not be obtained in complete generality. We will assume that there are two numbers $\omega, r < \infty$ such that

$$\mu : w \in \mathbb{R}^m \mapsto \frac{\omega}{2} \|w\|^2 + \sigma(w) \in \mathbb{R} \text{ is a convex mapping} \tag{1.11}$$

and such that it holds

$$\|w\|^2 \leq r(r - \sigma(w)) \quad \text{and} \quad -\sigma(w) \leq r(1 + \|w\|^2), \quad \forall w \in \mathbb{R}^m. \tag{1.12}$$

These two conditions are reasonable for the entropy σ . They amount to ask that the eigenvalues of its Hessian matrix (which are negative numbers because of concavity) are bounded from below and also bounded away from zero sufficiently far from the origin.

We now present the plan of our paper. In Section 2, we introduce the main notations that will be used throughout that article. In Section 3, we list some basic properties of the mappings S and g and of the integration formula $(\cdot, \cdot)_{0,h}$. In Section 4, we study the differential equation (1.5–1.7) and we prove a stability result. In Section 5, we define the numerical scheme we intend to study, we also explain how it can be implemented and we derive a stability result. We conclude in Section 6 by establishing convergence properties that are independent of the mesh size h .

2. NOTATIONS

We denote by $x \cdot y$ the Euclidean scalar product of $x, y \in \mathbb{R}^m$ and the Euclidean norm of x is denoted by $\|x\| = \sqrt{x \cdot x}$. We use the standard notation for Sobolev spaces, $L^2(\Omega)^m$ is the space of all the functions $v : \Omega \rightarrow \mathbb{R}^m$ such that $\|v\|_0^2 \equiv \int_{\Omega} \|v(x)\|^2 dx$ is finite and $H^1(\Omega)^m = \{v \in L^2(\Omega)^m \mid \partial_1 v, \partial_2 v, \partial_3 v \in L^2(\Omega)^m\}$. We equip $H^1(\Omega)^m$ with the norm $\|v\|_1^2 \equiv \|v\|_0^2 + |v|_1^2$ where $|v|_1^2 \equiv \|\partial_1 v\|_0^2 + \|\partial_2 v\|_0^2 + \|\partial_3 v\|_0^2$. We define $H_0^1(\Omega)^m$ as the closure of $C_0^\infty(\Omega)^m$ in $H^1(\Omega)^m$ and $H^{-1}(\Omega)^m$ as the dual space to $H^1(\Omega)^m$. The notations $(\cdot, \cdot)_0$ and $(\cdot, \cdot)_1$ will be used for the standard scalar product in $L^2(\Omega)^m$ and in $H_0^1(\Omega)^m$, $(u, v)_0 = \int_{\Omega} u(x) \cdot v(x) dx$ and $(u, v)_1 = \sum_{i=1}^3 (\partial_i u, \partial_i v)_0$. If $T > 0$ and if Z is a Banach space with norm $\|\cdot\|_Z$, $L^2(0, T; Z)$ will be the space of all the functions $v : (0, T) \rightarrow Z$ such that $\|v\|_{L^2(0,T;Z)}^2 \equiv \int_{\Omega} \|v(t)\|_Z^2 dt < \infty$.

We conclude with some notations relative to the discrete spaces V_h and W_h of piecewise linear and piecewise constant functions which have been introduced in Section 1. We use the integration formula $(\cdot, \cdot)_{0,h}$ (see (1.8)) to construct two mesh depending norms $\|\cdot\|_{-1,h}$ and $\|\cdot\|_{*,h}$,

$$\|v\|_{-1,h} \equiv \sup_{\substack{\psi \in V_h \\ \psi \neq 0}} \frac{(v, \psi)_{0,h}}{|\psi|_1}, \quad \|v\|_{*,h} \equiv \sup_{\substack{\psi \in V_h \\ \psi \neq 0}} \frac{(v, \psi)_0}{|\psi|_1}, \quad v \in W_h. \tag{2.1}$$

We will finally denote by $L^2(0, T; H_h^{-1}(\Omega))$ the space of all the functions $v : (0, T) \rightarrow W_h$ such that the quantity $\|v\|_{L^2(0, T; H_h^{-1}(\Omega))}^2 \equiv \int_0^T \|v(t)\|_{-1, h}^2 dt$ is finite.

3. BASIC PROPERTIES OF THE SEMI-DISCRETE PROBLEM (1.5–1.7)

In this section, we will establish some basic properties of the bilinear form $(\cdot, \cdot)_{0, h}$ defined in (1.8) and of the functions S and g introduced in (1.9). These properties are essential to analyze the efficiency of the implicit integration scheme we will propose later on. The first one concerns the entropy S . To prove it, we will need the following Lemma.

Lemma 3.1. *Let $\lambda \in C^1(\mathbb{R}^m; \mathbb{R})$ be a concave function and assume that $\omega > 0$ is so large that $\mu : x \rightarrow \frac{\omega}{2}\|x\|^2 + \lambda(x)$ is convex. Then, for any $x, y \in \mathbb{R}^m$, it holds*

$$\omega\|y - x\|^2 \geq -(\nabla\lambda(y) - \nabla\lambda(x))^T(y - x) \geq \frac{1}{\omega}\|\nabla\lambda(y) - \nabla\lambda(x)\|^2. \tag{3.1}$$

Proof. Because of the assumptions made on λ and μ the function $f(s) = \mu(x + s(y - x))$ is differentiable and convex. It thus holds

$$f'(1) - f'(0) \geq 0. \tag{3.2}$$

By the chain rule we have $f'(1) = \nabla\mu(y)^T(y - x)$ and $f'(0) = \nabla\mu(x)^T(y - x)$. It thus follows from the definition of μ that $f'(1) = (\omega y + \nabla\lambda(y))^T(y - x)$ and that $f'(0) = (\omega x + \nabla\lambda(x))^T(y - x)$. We substitute these relations in (3.2) and we get the left-hand side of (3.1).

To prove the right-hand side of (3.1), we proceed in two steps. We first suppose that λ is of class C^2 . In that case, μ is also of class C^2 and we denote by H_λ and H_μ the Hessian matrices to λ and μ . We choose $x, y \in \mathbb{R}^m$ and we set

$$J_\lambda = - \int_0^1 H_\lambda(x + s(y - x)) ds \quad \text{and} \quad J_\mu = \int_0^1 H_\mu(x + s(y - x)) ds. \tag{3.3}$$

Since λ is concave and μ convex, J_λ and J_μ are symmetric positive semi-definite. Moreover, the definition of μ implies that $H_\mu(w) = \omega\mathbb{I} + H_\lambda(w)$. Substituting this relation in the definition for J_μ we get that $J_\mu = \omega\mathbb{I} - J_\lambda$ which proves that J_μ and J_λ commute. The conclusion is that the product $J_\mu J_\lambda$ is also a symmetric positive semi-definite matrix. In particular

$$(J_\mu(x - y))^T(J_\lambda(x - y)) \geq 0. \tag{3.4}$$

Taking the definition (3.3) of J_λ and J_μ into account, we deduce from the fundamental Theorem of Analysis that $J_\mu(x - y) = \nabla\mu(x) - \nabla\mu(y)$ and that $J_\lambda(x - y) = -(\nabla\lambda(x) - \nabla\lambda(y))$. Therefore (3.4) can be rewritten as

$$-(\nabla\mu(x) - \nabla\mu(y))^T(\nabla\lambda(x) - \nabla\lambda(y)) \geq 0$$

which proves the right-hand side of (3.1), because $\nabla\mu(w) = \omega w + \nabla\lambda(w)$ by definition of μ .

If λ fails to be of class C^2 , we introduce a family of mollifiers $\{\rho_n\} \subset C^\infty(\mathbb{R}^m; \mathbb{R}_+)$ satisfying the classical conditions $\text{supp}(\rho_n) \subset B(0, \frac{1}{n})$, $\int_{\mathbb{R}^m} \rho_n(w) dw = 1$ and we consider the sequences $\{\lambda_n\}$ and $\{\mu_n\}$,

$$\lambda_n(x) = \int_{\mathbb{R}^m} \lambda(x + z)\rho_n(z) dz, \quad \mu_n(x) = \int_{\mathbb{R}^m} \mu(x + z)\rho_n(z) dz, \quad x \in \mathbb{R}^m. \tag{3.5}$$

Since λ is concave and μ convex, one easily proves that λ_n is concave and μ_n convex. Moreover, the definition of μ and easy computations show that $\mu_n(x)$ only differs from $\frac{\omega}{2}\|x\|^2 + \lambda_n(x)$ by the addition of an affine function. The convexity of μ_n thus implies that $x \rightarrow \frac{\omega}{2}\|x\|^2 + \lambda_n(x)$ is also a convex function and the right-hand side of inequality (3.1) is valid for the regular mapping $\lambda_n \in C^\infty(\mathbb{R}^m, \mathbb{R})$ (see Prop. IV.20 in [5]). It holds

$$-(\nabla\lambda_n(y) - \nabla\lambda_n(x))^T(y - x) \geq \frac{1}{\omega}\|\nabla\lambda_n(y) - \nabla\lambda_n(x)\|^2, x, y \in \mathbb{R}^m.$$

Letting n tend to ∞ , we conclude that (3.1) is also true for λ because $\nabla\lambda_n$ converge uniformly to $\nabla\lambda$ on each compact subset of \mathbb{R}^m (see Lem. IX.1 and Prop. IV.21 in [5]). \square

We are now in a position to state the central properties of the entropy functional S and of the source term g .

Theorem 3.1. (a) *Assume that σ is a C^1 and concave mapping defined on \mathbb{R}^m and that it satisfies conditions (1.11) and (1.12) for some numbers r and ω . Then the functional S defined in (1.9) satisfies*

$$\|u\|_0^2 \leq r(r - S(u)), \quad -S(u) \leq r(1 + \|u\|_0^2), \quad u \in L^2(\Omega)^m, \tag{3.6}$$

$$\omega\|u - v\|_0^2 \geq -(\partial S(u) - \partial S(v), u - v)_0 \geq \frac{1}{\omega}\|\partial S(u) - \partial S(v)\|_0^2, \quad u, v \in L^2(\Omega)^m \tag{3.7}$$

$$-(\partial S(u), v - u)_0 \geq -S(v) + S(u) - \frac{\omega}{2}\|u - v\|_0^2, \quad u, v \in L^2(\Omega)^m, \tag{3.8}$$

$$\omega\|u - v\|_0 \geq \|\partial S(u) - \partial S(v)\|_0, \quad u, v \in L^2(\Omega)^m. \tag{3.9}$$

(b) *If the mapping $G : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is Lipschitz continuous and fulfils condition (1.3), the function g defined in (1.9) satisfies*

$$(\partial S(u), g(u))_0 \geq 0, \quad u \in L^2(\Omega)^m \tag{3.10}$$

and there is $C < \infty$ independent of h such that it holds

$$\|g(u) - g(v)\|_0 \leq C\|\partial S(v) - \partial S(u)\|_0, \quad u, v \in L^2(\Omega)^m, \tag{3.11}$$

$$\|g(u)\|_0 \leq C(1 + \|u\|_0), \quad u \in L^2(\Omega)^m. \tag{3.12}$$

Proof. We proceed in five steps. In the first step, we prove (3.6), in the second we establish (3.7), (3.10) and (3.11). The relation (3.9) will be established in the third step while (3.8) and (3.12) will be proved during the two last steps.

First step. The relation (3.6) is a direct consequence of the definition (1.9) of S and of the property (1.12) of σ .

Second step. Using (1.9) to compute the differential ∂S to S , one gets that

$$\partial S(v)(x) = \nabla\sigma(v(x)), \quad x \in \Omega, \quad v \in L^2(\Omega)^m. \tag{3.13}$$

This remark helps us to prove (3.7) as well as (3.10) and (3.11). Let $u, v \in L^2(\Omega)^m$. Since the entropy σ is a C^1 concave function satisfying (1.11), Lemma 3.1 and (3.13) imply that

$$\omega\|u(x) - v(x)\|^2 \geq -(\partial S(u)(x) - \partial S(v)(x))^T(u(x) - v(x)) \geq \frac{1}{\omega}\|\partial S(u)(x) - \partial S(v)(x)\|^2,$$

for any $x \in \Omega$. Taking the definition (1.9) of g into account, we deduce from (3.13) and the second principle of thermodynamics (see (1.3)) that

$$g(u)(x) \cdot \partial S(u)(x) = G(\nabla\sigma(u(x))) \cdot \nabla\sigma(u(x)) \geq 0, \quad x \in \Omega.$$

Finally, denoting by L_G the Lipschitz constant of the source term G , we deduce from (3.13) that any $x \in \Omega$ satisfies

$$\|g(u)(x) - g(v)(x)\|^2 = \|G(\nabla\sigma(u(x))) - G(\nabla\sigma(v(x)))\|^2 \leq L_G \|\partial S(u)(x) - \partial S(v)(x)\|^2.$$

Integrating the three above relations over Ω gives (3.7), (3.10) and (3.11).

Third step. Because of Cauchy-Schwarz inequality, it follows from the right-hand side of (3.7) that

$$\|\partial S(u) - \partial S(v)\|_0 \|u - v\|_0 \geq \frac{1}{\omega} \|\partial S(u) - \partial S(v)\|_0^2$$

and dividing this relation by $\|\partial S(u_2) - \partial S(u_1)\|_0$ establishes the Lipschitz condition (3.9).

Fourth step. Starting from the definition of the differential ∂S , we write the identity

$$S(u_2) - S(u_1) = (\partial S(u_1), u_2 - u_1) + \int_0^1 (\partial S(u(s)) - \partial S(u_1), u_2 - u_1) ds \tag{3.14}$$

where $u(s) = u_1 + s(u_2 - u_1)$, $s \in (0, 1)$. Since $u_2 - u_1 = \frac{1}{s}(u(s) - u_1)$, the right-hand side of (3.7) implies that $(\partial S(u(s)) - \partial S(u_1), u_2 - u_1) \geq -\frac{\omega}{s} \|u(s) - u_1\|_0^2 = -\omega s \|u_2 - u_1\|_0^2$. We substitute that estimate in (3.14) and get that

$$S(u_2) - S(u_1) \geq (\partial S(u_1), u_2 - u_1) - \omega \|u_2 - u_1\|_0^2 \int_0^1 s ds,$$

which proves (3.8).

Last step. We use (3.11), the triangle inequality and the Lipschitz property (3.9) of S to write that

$$\|g(u)\|_0 \leq \|g(0)\|_0 + C \|\partial S(u) - \partial S(0)\|_0 \leq \|g(0)\|_0 + C\omega \|u\|_0.$$

Since it is clear from its definition (1.9) that $\|g(0)\|_0 \leq |\Omega|^{\frac{1}{2}} \|G(\nabla\sigma(0))\|$, the relation (3.12) follows. □

The next Lemma is related to the mesh depending norms $\|\cdot\|_{-1,h}$ and $\|\cdot\|_{*,h}$ defined in (2.1).

Lemma 3.2. *Assume that the meshes \mathcal{M}_h are regular and satisfy an inverse condition. Then there is a constant c independent of h such that*

$$\|v\|_{*,h} \leq \|v\|_{-1,h} + ch \|v\|_0, \quad \|v\|_{-1,h} \leq \|v\|_{*,h} + ch \|v\|_0, \quad v \in W_h, \tag{3.15}$$

$$h \|v\|_0 \leq c \|v\|_{-1,h}, \quad v \in W_h. \tag{3.16}$$

Proof. From the definition of the mesh size h it follows that $\min_{1 \leq k \leq N} \text{dist}(P, P_k) \leq h$ for any point $P \in \Omega$. Consequently, each Delaunay cell \mathcal{O}_j is necessarily contained in the ball of radius h centred in P_j and a scaling argument implies that there is $C < \infty$, independent of h , such that

$$\int_{\mathcal{O}_j} \|\psi(x) - \psi(P_j)\|^2 dx \leq Ch^2 \int_{\mathcal{O}_j} \|\nabla\psi(x)\|^2 dx, \quad j = 1, 2, \dots, N, \quad \psi \in V_h. \tag{3.17}$$

Since the Delaunay cells exactly cover Ω , summing (3.17) over $j = 1, 2 \dots N$ establishes the error estimate

$$\|\psi - I_h\psi\|_0 \leq ch|\psi|_1, \psi \in V_h. \tag{3.18}$$

where I_h is the interpolant operator onto W_h . For any $\psi \in V_h$, $I_h\psi$ is the unique element of W_h satisfying $I_h\psi|_{\mathcal{O}_j} = \psi(P_j)$, $j = 1, 2 \dots N$. Observe that I_h may be used to rewrite the definition (1.8) of the integration formula as $(v, \psi)_{0,h} = (v, I_h\psi)_0$, $v \in W_h$, $\psi \in V_h$. Hence (3.18) and the Cauchy-Schwarz inequality imply that $|(v, \psi)_0 - (v, \psi)_{0,h}| \leq ch\|v\|_0|\psi|_1$ for any $v \in W_h$, $\psi \in V_h$. From the triangle inequality, we deduce that

$$\frac{(v, \psi)_0}{|\psi|_1} \leq \frac{(v, \psi)_{0,h}}{|\psi|_1} + ch\|v\|_0, \quad \frac{(v, \psi)_{0,h}}{|\psi|_1} \leq \frac{(v, \psi)_0}{|\psi|_1} + ch\|v\|_0, \quad v \in W_h, \psi \in V_h, \psi \neq 0$$

and the thesis (3.15) follows from the definition (2.1) of the mesh depending norms $\|\cdot\|_{-1,h}$ and $\|\cdot\|_{*,h}$.

We now prove (3.16). We take $v \in W_h$ and construct $\psi \in V_h$ by prescribing its value at any node P_j , $\psi(P_j) = v(P_j)$, $j = 1 \dots N$. Because of the definition (1.8) of $(\cdot, \cdot)_{0,h}$ and since v is piecewise constant on the Delaunay cells, we have the identity

$$(v, \psi)_{0,h} = \sum_{j=1}^{j=N} |\mathcal{O}_j| \|v(P_j)\|^2 = \|v\|_0^2. \tag{3.19}$$

On the other hand, we may also write that

$$(v, \psi)_{0,h} = \sum_{j=1}^{j=N} |\mathcal{O}_j| \|\psi(P_j)\|^2. \tag{3.20}$$

Let us denote by T_j the union of all the tetrahedrons in \mathcal{M}_h sharing the node P_j . Since the meshes are regular and as they satisfy an inverse condition, there is $C_r < \infty$, independent of h , such that $C_r^2|\mathcal{O}_j| \geq |T_j|$, $j = 1 \dots N$ and we deduce from (3.20) that $C_r^2(v, \psi)_{0,h} \geq \sum_{j=1}^{j=N} |T_j| \|\psi(P_j)\|^2$. Observe that the right-hand side of this inequality is larger than $\|\psi\|_0^2$. Hence, it holds $(v, \psi)_{0,h} \geq C_r^{-2}\|\psi\|_0^2$ and combining this relation with (3.19) yields

$$C_r^{-1}\|v\|_0\|\psi\|_0 \leq (v, \psi)_{0,h}. \tag{3.21}$$

Recall now that \mathcal{M}_h satisfies an inverse condition and that $\psi \in V_h$. It thus holds $C_i h|\psi|_1 \leq \|\psi\|_0$ for some C_i independent of h [19]. We substitute that information in (3.21). We get that $C_r^{-1}C_i h\|v\|_0 \leq \frac{(v, \psi)_{0,h}}{|\psi|_1}$ and (3.16) follows because $\frac{(v, \psi)_{0,h}}{|\psi|_1} \leq \|v\|_{-1,h}$ (see (2.1)). □

We now show that the standard norm of $H^{-1}(\Omega)^m$ is dominated by the mesh depending norm $\|\cdot\|_{-1,h}$.

Theorem 3.2. *Assume that the meshes \mathcal{M}_h are regular and satisfy an inverse condition. Then there is $b_1, b_2 > 0$ independent of h such that*

$$b_1\|v\|_{-1} \leq \|v\|_{-1,h} \leq b_2\|v\|_0, v \in W_h. \tag{3.22}$$

Proof. Take $v \in W_h$. Since the Laplace operator is an isomorphism from $H_0^1(\Omega)^m$ into $H^{-1}(\Omega)^m$, it holds $\|v\|_{-1} = |\chi|_1$ where $\chi \in H_0^1(\Omega)^m$ satisfies $-\Delta\chi = v$.

It is an easy exercise to deduce from the definition (2.1) that $\|v\|_{*,h} = |\Pi_h\chi|_1$ if Π_h is the $H_0^1(\Omega)$ -orthogonal projector onto V_h . It thus follows that $\|v\|_{-1} - \|v\|_{*,h} = |\chi|_1 - |\Pi_h\chi|_1$ and one can use the inverse triangle

inequality to bound the left-hand side. It yields

$$\|v\|_{-1} \leq \|v\|_{*,h} + |\chi - \Pi_h \chi|_1. \tag{3.23}$$

Since Ω is a convex polyhedron, the pre-Laplacian χ to $v \in W_h \subset L^2(\Omega)^m$ belongs to $H^2(\Omega)^m$ [17]. It follows from the standard approximation properties of finite element spaces [19] that $|\chi - \Pi_h \chi|_1 \leq C_a h \|\chi\|_2$ for some C_a independent of h . Using that the shift inequality $\|\chi\|_2 \leq C_s \|\Delta \chi\|_0$ is valid for any $\chi \in H^2(\Omega)^m \cap H_0^1(\Omega)^m$ [17], one concludes that $|\chi - \Pi_h \chi|_1 \leq C_a C_s h \|v\|_0$. We substitute that information into (3.23) and we also use the first statement in (3.15) which provides an estimate of $\|v\|_{*,h}$ in terms of $\|v\|_{-1,h}$. We finally get that

$$\|v\|_{-1} \leq \|v\|_{-1,h} + Ch \|v\|_0 \tag{3.24}$$

where $C < \infty$ does not depend on h . Taking into account that $v \in W_h$, we are allowed to use the inverse inequality (3.16). This operation achieves the proof of the first part of (3.22).

By the Cauchy-Schwarz inequality and the Sobolev injection of $H_0^1(\Omega)^m$ into $L^2(\Omega)^m$ we have $(v, \psi)_0 \leq C \|v\|_0 |\psi|_1$ for any $v \in W_h, \psi \in V_h$ and where C only depends on Ω . This inequality and the definition (2.1) imply that $\|v\|_{*,h} \leq C \|v\|_0$ for any $v \in W_h$. The left-hand side of (3.22) follows from that last relation and the estimate for $\|v\|_{-1,h}$ established in (3.15). \square

From now on, we will never refer to the exact definitions (1.9) and (1.8) of the mappings S, g and of the form $(\cdot, \cdot)_{0,h}$. All the stability and convergences results will only be based on the properties stated in Theorems 3.1 and 3.2 and on the fact that the $L^2(\Omega)^m$ norm of the initial data u^0 (see (1.7)) is bounded independently of h ,

$$\|u^0\|_0 \leq \|w^0\|_0, \tag{3.25}$$

because it is the $L^2(\Omega)^m$ -projection of w^0 onto W_h .

To conclude this section, we use the Riesz Theorem [11] and the fact that $(\cdot, \cdot)_1$ and $(\cdot, \cdot)_0$ are scalar products and define the two $B \in \mathcal{L}(W_h, V_h)$ and $B^* \in \mathcal{L}(V_h, W_h)$:

$$\begin{cases} \forall v \in W_h, Bv \text{ is the only element in } V_h \text{ s.t. } (Bv, \phi)_1 = (v, \phi)_{0,h}, \phi \in V_h, \\ \forall \psi \in V_h, B^*\psi \text{ is the only element in } W_h \text{ s.t. } (B^*\psi, w)_0 = (w, \psi)_{0,h}, w \in W_h. \end{cases} \tag{3.26}$$

Remark 3.1. The operators B and B^* are useful to rewrite the system (1.5–1.6) in a more compact form:

$$B\dot{u}(t) = Bg(u(t)) + \phi(t), \quad \forall t \in (0, T), \tag{3.27}$$

$$B^*\phi(t) = \partial S(u(t)), \quad \forall t \in [0, T]. \tag{3.28}$$

Remark 3.2. The result of Theorem 3.2 can also be reformulated in terms of B or B^* . It follows from (3.26) and from the definition (2.1) of $\|\cdot\|_{-1,h}$ that

$$|Bv|_1 = \|v\|_{-1,h}, \quad v \in W_h. \tag{3.29}$$

We thus may rewrite (3.22) as

$$b_1 \|v\|_{-1} \leq |Bv|_1 \leq b_2 \|v\|_0, \quad v \in W_h. \tag{3.30}$$

Remark 3.3. The composed operator B^*B is a linear transformation of the space W_h . Using the identity $(B^*Bv, v)_0 = |Bv|_1^2$, which is a direct consequence of definitions (3.26), one can deduce from (3.30) that B^*B is bijective. This properties finally implies that B^* is a linear bijection from the Range of B into W_h .

4. EXISTENCE AND UNIQUENESS FOR THE SOLUTION TO (1.5–1.7)

In this section we prove the following theorem.

Theorem 4.1. *Under the assumptions of Theorem 3.1 and 3.2, the problem (1.5–1.7) has a unique solution (u, ϕ) in $C^1([0, T]; W_h) \times C^0([0, T]; V_h)$ and it holds*

$$\|\dot{u}\|_{L^2(0,T;H_h^{-1}(\Omega))} + \|u\|_{L^\infty(0,T;L^2(\Omega))} + \|\phi\|_{L^2(0,T;H_0^1(\Omega))} \leq c, \tag{4.1}$$

for a constant c independent on the mesh size h .

Proof. To prove existence and uniqueness we use the mapping B^*B defined in Remarks 3.2 and 3.3. Since this operator is bijective and as ∂S and g are assumed to be Lipschitz continuous from W_h into W_h (see (3.9) and (3.11) in Theo. 3.1), the Cauchy-Lipschitz Theorem [14] makes sure that there exists exactly one $u \in C^1([0, T]; W_h)$ satisfying the initial condition (1.7) and the ordinary differential equation

$$B^*B\dot{u}(t) = B^*Bg(u(t)) + \partial S(u(t)), t \in [0, T]. \tag{4.2}$$

Since B^* is a bijection from the Range of B to W_h (see Rem. 3.3), we define $\phi \in C^0([0, T]; \text{Range } B)$ by asking that the couple (u, ϕ) satisfies (3.28). Substituting this relation into (4.2), we get that $B^*B\dot{u}(t) = B^*Bg(u(t)) + B^*\phi(t)$ which means that (u, ϕ) also fulfills (3.27) because B^* is injective on the Range of B (see Rem. 3.3). As equations (3.27) and (3.28) are equivalent to (1.5) and (1.6), we have proved that the couple (u, ϕ) is a solution to (1.5–1.7). The uniqueness result is clear. Assume that (u_1, ϕ_1) and (u_2, ϕ_2) are solutions to (3.27), (3.28) and (1.7). Multiplying (3.27) by B^* and taking (3.28) into account, we get that u_1 and u_2 are two solutions to o.d.e. (4.2) satisfying the same initial condition (1.7). It thus follows from the Cauchy-Lipschitz Theorem that

$$u_1 = u_2. \tag{4.3}$$

Because of (3.27), $\phi_1(t), \phi_2(t) \in \text{Range } B, t \in [0, T]$. Since B^* is injective on $\text{Range } B$ (see Rem. 3.3), (3.28) and (4.3) together imply that $\phi_1 = \phi_2$ which proves uniqueness.

It remains to show that the pair (u, ϕ) solving (1.5–1.7) satisfies (4.1). We take $\xi = \phi(t)$ as a test function in (1.5) and $v = \dot{u}(t)$ as a test function in (1.6). Combining the results, we get that

$$|\phi(t)|_1^2 - (\partial S(u(t)), \dot{u}(t))_0 = -(g(u(t)), \phi(t))_{0,h}, t \in (0, T). \tag{4.4}$$

Testing (1.6) against $v = g(u(t))$ and taking the property (3.10) into account, we observe that the right-hand side of (4.4) is non-positive and it holds $|\phi(t)|_1^2 - (\partial S(u(t)), \dot{u}(t))_0 \leq 0$. Integrating that relation and using the initial condition (1.7), we get that

$$\|\phi\|_{L^2(0,t;H_0^1(\Omega))}^2 - S(u(t)) \leq -S(u^0), t \in [0, T].$$

We add the number r on both side and we use that $r(r - S(u(t))) \geq \|u(t)\|_0^2$ and that $-S(u^0) \leq r(1 + \|u^0\|_0^2)$ (see (3.6)). We find that

$$\|\phi\|_{L^2(0,t;H_0^1(\Omega))}^2 + \frac{1}{r}\|u(t)\|_0^2 = 2r + r\|u^0\|_0^2, \forall t \in [0, T]$$

and, as $\|u^0\|_0^2$ is bounded independently on h (see (3.25)), we have proved that

$$\|u\|_{L^\infty(0,T;L^2(\Omega))} + \|\phi\|_{L^2(0,T;H_0^1(\Omega))} \leq c \tag{4.5}$$

for some constant c which does not depend on h . To achieve the proof we use the differential equation (1.5) and the definition (2.1) for the norm $\|\cdot\|_{-1,h}$ to observe that $\|\dot{u}(t)\|_{-1,h} \leq |\phi(t)|_1 + \|g(u(t))\|_{-1,h}$, $t \in (0, T)$. Because of the properties (3.22) and (3.12), the last term in the right-hand side turns out to be less than $b_2L(1 + \|u(t)\|_0)$. It thus holds

$$\|\dot{u}(t)\|_{-1,h} \leq |\phi(t)|_1 + b_2L(1 + \|u(t)\|_0), \quad t \in (0, T). \tag{4.6}$$

and the full estimate (4.1) follows from (4.6) and the partial estimate (4.5). □

5. DEFINITION OF AN IMPLICIT SCHEME TO SOLVE PROBLEM (1.5–1.7)

Let $\tau > 0$ be a time step and denote by M the largest integer such that $M\tau \leq T$. For any $i = 1, 2 \dots M$, we are looking for a pair $(u_i, \phi_i) \in W_h \times V_h$ approaching the values $(u(t_i), \phi(t_i))$ of the solution (u, ϕ) to (1.5–1.7) at time $t_i = i\tau$. Our idea is to construct (u_i, ϕ_i) in an inductive way. We start from the initial condition

$$u_0 = u^0. \tag{5.1}$$

Then, u_{i-1} being given, we compute (u_i, ϕ_i) as the solution to

$$(u_i - u_{i-1}, \xi)_{0,h} - \tau(\phi_i, \xi)_1 = \tau(g(u_{i-1}), \xi)_{0,h}, \quad i = 1, 2 \dots M, \quad \xi \in V_h, \tag{5.2}$$

$$(v, \phi_i)_{0,h} - (\partial S(u_{i-1}) - \beta(u_i - u_{i-1}), v)_0 = 0, \quad i = 1, 2 \dots M, \quad v \in W_h, \tag{5.3}$$

where $\beta \geq 0$ is a stabilization parameter. At first we check that the system (5.2–5.3) is well posed.

Theorem 5.1. *Under the assumptions of Theorem 3.2 there is exactly one sequence (u_i, ϕ_i) , $i = 1, 2 \dots M$ solution to (5.2) and (5.3).*

Proof. Recall that W_h and V_h are finite dimensional spaces, (5.2) and (5.3) may thus be seen as a linear system with $\dim W_h + \dim V_h$ unknowns (the components of u_i and ϕ_i) and with the same number of equations. Because of the Fredholm alternative, the entire Theorem will be proved if we show that (5.2) and (5.3) has at most one solution (u_i, ϕ_i) .

The difference U and Φ between two possible solution is such that

$$(U, \xi)_{0,h} - \tau(\Phi, \xi)_1 = 0, \quad \xi \in V_h, \tag{5.4}$$

$$(v, \Phi)_{0,h} + \beta(U, v)_0 = 0, \quad v \in W_h. \tag{5.5}$$

We take $\xi = \Phi$ and $v = U$ as test functions in (5.4) and in (5.5). We get that $(U, \Phi)_{0,h} = \tau\|\Phi\|_1^2 = -\beta\|U\|_0^2$. Since β is non-negative and τ positive, this relation implies that $\Phi = 0$. It then follows from (5.4) that $(U, \xi)_{0,h} = 0$ for any $\xi \in V_h$ which means that $\|U\|_{-1,h} = 0$, i.e that $U = 0$, because of (3.22). The Theorem is proved. □

5.1. An implementation of the scheme (5.1–5.3).

The scheme (5.1–5.3) has been implemented in [16]. Let χ_j be the characteristic function of the Delaunay cells \mathcal{O}_j , $1 \leq j \leq N$, and let e_k , $1 \leq k \leq M$, be the canonical vectors spanning \mathbb{R}^m . The family $\{\chi_j e_k\}$ is a basis of the space W_h containing the \mathbb{R}^m -valued functions that are piecewise constant on \mathcal{M}'_h . We can observe that

- (i) the mass matrix $D_{ikjk'} = (\chi_i e_k, \chi_j e_{k'})_0$ is diagonal,
- (ii) the vectors \mathbf{g}^{i-1} and \mathbf{F}^{i-1} containing the components of $g(u_{i-1}) \in W_h$ and $\partial S(u_{i-1}) \in W_h$ onto the basis $\{\chi_j e_k\}$ are easy to compute when the corresponding components \mathbf{u}^{i-1} to $u_{i-1} \in W_h$ are known.

Let ψ_j be the standard \mathbb{P}^1 -hat functions associated to the nodes $P_j, j = 1 \dots N$. The family $\{\psi_j e_k\}$ spans the space V_h containing the \mathbb{R}^m -valued functions that are piecewise linear on \mathcal{M}_h . We define the rigidity matrix $A_{ikjk'} = (\psi_i e_k, \psi_j e_{k'})_0$ and the rectangular matrix $M_{ikjk'} = (\chi_j e_k, \psi_i e_{k'})_{0,h}$. With this material, we are in a position to rewrite equations (5.2) and (5.3) as a linear system for the components \mathbf{u}^i and ϕ^i to u_i and ϕ^i :

$$M\mathbf{u}^i - \tau A\phi^i = M\mathbf{u}^{i-1} + \tau M\mathbf{g}^{i-1} \tag{5.6}$$

$$\beta D\mathbf{u}^i + M^T \phi^i = \beta D\mathbf{u}^{i-1} + D\mathbf{F}^{i-1}. \tag{5.7}$$

Since D is diagonal, \mathbf{u}^i can be eliminated from (5.7),

$$\mathbf{u}^i = \mathbf{u}^{i-1} + \frac{1}{\beta} (\mathbf{F}^{i-1} - D^{-1} M^T \phi^i). \tag{5.8}$$

We substitute this relation in (5.6) and we obtain a linear equation for ϕ^i alone,

$$\left(\tau A + \frac{1}{\beta} M D^{-1} M^T \right) \phi^i = \tau M\mathbf{g}^{i-1} + \frac{1}{\beta} M\mathbf{F}^{i-1}. \tag{5.9}$$

The linear system (5.9) can be solved by the Choleski's method or by the gradient conjugate algorithm because the governing matrix is symmetric positive definite. The components \mathbf{u}^i to u_i are finally recovered thanks to relation (5.8).

5.2. Estimates for the solution to (5.1–5.3).

We prove an estimate for the solution (u_i, ϕ_i) to (5.1–5.3) which is uniformly valid with respect to h under a stability constraint. This condition is based on the parameters ω and b_1 introduced in Theorems 3.1 and 3.2 and on the constant ρ in the inverse inequality,

$$\|v\|_0 \leq \rho b_1 \|v\|_{-1}, v \in W_h. \tag{5.10}$$

It reads as follows: There must be $\alpha, \varepsilon > 0$ such the stabilization parameter β and the time step τ satisfy

$$\left(\frac{\omega}{2} - \beta + \alpha \right) \rho^2 \tau \leq 1 - 2\varepsilon. \tag{5.11}$$

Theorem 5.2. *We assume that the assumptions in Theorems 3.1 and 3.2 are fulfilled and that condition (5.11) is valid for some $\alpha, \varepsilon > 0$. Then the sequence (u_i, ϕ_i) , solution to (5.1–5.3), satisfies*

$$\max_{0 \leq i \leq M} \|u_i\|_0^2 + \sum_{i=1}^{i=M} \|u_i - u_{i-1}\|_0^2 + \tau \sum_{i=1}^{i=M} |\phi_i|_1^2 \leq c. \tag{5.12}$$

for a constant c independent of h and of τ .

Proof. We take $\xi = \phi_i$ and $v = u_i - u_{i-1}$ as test functions in (5.1) and in (5.3) respectively. We use property (3.8) to estimate $-(\partial S(u_{i-1}), u_i - u_{i-1})$ and we bound $|(g(u_{i-1}, \phi_i)_{0,h})|$ by mean of $\|g(u_{i-1})\|_{-1,h} |\phi_i|_1$ (see the definition (2.1) of $\|g(u_{i-1})\|_{-1,h}$). As a conclusion we get, for $i = 1, 2 \dots M$,

$$-S(u_i) + \alpha \|u_i - u_{i-1}\|_0^2 + \tau |\phi_i|_1^2 \leq -S(u_{i-1}) + \tau \|g(u_{i-1})\|_{-1,h} |\phi_i|_1 + \left(\frac{\omega}{2} - \beta + \alpha \right) \|u_i - u_{i-1}\|_0^2. \tag{5.13}$$

We need a bound for the last term in the right-hand side. The equation (5.2) for the difference $u_i - u_{i-1}$ imply that

$$\|u_i - u_{i-1}\|_{-1,h} \leq \tau (|\phi_i|_1 + \|g(u_{i-1})\|_{-1,h}).$$

Because of (3.22), the left-hand side of this inequality is larger than $b_1 \|u_i - u_{i-1}\|_{-1}$ and it follows from the inverse inequality (5.10) that

$$\|u_i - u_{i-1}\|_0^2 \leq \rho^2 \tau (\tau |\phi_i|_1^2 + 2\tau \|g(u_{i-1})\|_{-1,h} |\phi_i|_1 + \tau \|g(u_{i-1})\|_{-1,h}^2).$$

Combining that estimate with the stability condition (5.11), we deduce from (5.13) that

$$-S(u_i) + \alpha \|u_i - u_{i-1}\|_0^2 + 2\varepsilon \tau |\phi_i|_1^2 \leq -S(u_{i-1}) + 3\tau \|g(u_{i-1})\|_{-1,h} |\phi_i|_1 + \tau \|g(u_{i-1})\|_{-1,h}^2.$$

We apply the Young's inequality to the product $\|g(u_{i-1})\|_{-1} |\phi_i|_1$ and we combine (3.6), (3.22) and (3.12) to bound $\|g(u_{i-1})\|_{-1,h}^2$ by $r(r - S(u_{i-1}))$. We obtain

$$-S(u_i) + \alpha \|u_{i-1} - u_i\|_0^2 + \varepsilon \tau |\phi_i|_1^2 \leq -(1 + \tau C)S(u_{i-1}) + \tau C, \quad i = 1, 2 \dots M.$$

for some constant $C < \infty$ independent of h and τ . We use the discrete Gronwall Lemma and we take the initial condition (5.1) into account. We conclude that the constant C can be chosen large enough so that it holds

$$\max_{0 \leq i \leq M} [-S(u_i)] + \sum_{i=1}^{i=M} \|u_{i-1} - u_i\|_0^2 + \tau \sum_{i=1}^{i=M} |\phi_i|_1^2 \leq C(C - S(u^0)).$$

To achieve the proof of the Theorem, we use (3.6) and (3.25) to bound $-S(u^0)$ independently of h and to estimate $\max_{0 \leq i \leq M} \|u_i\|_0^2$ by mean of $\max_{0 \leq i \leq M} [-S(u_i)]$. □

Remark 5.1. Observe that if the stabilization parameter β is larger than $\frac{\omega}{2}$ then the constraint (5.11) is fulfilled by any time step τ with the choice $\alpha = \beta - \frac{\omega}{2}$ and $\varepsilon = \frac{1}{2}$. In that case, the scheme (5.1–5.3) is said to be unconditionally stable.

6. CONVERGENCE RESULTS FOR THE SCHEME (5.1–5.3).

We now analyse the convergence property of the scheme (5.1–5.2). Let us observe that a higher convergence order can be obtained under the following property.

Property 6.1. *It holds $-(\partial S(u_2) - \partial S(u_1), u_2 - u_1)_0 \geq \nu \|u_2 - u_1\|_0^2$ for any $u_1, u_2 \in W_h$ and for some $\nu > 0$.*

Because of the definition (1.9) of S , property 6.1 amounts to ask that the eigenvalues to the Hessian matrix of the entropy $\sigma : \mathbb{R}^m \rightarrow \mathbb{R}$ are uniformly bounded away from zero. This condition is unfortunately not fulfilled in general. Thermodynamical systems undergoing phase transition are actually characterized by degenerate entropies having non-definite Hessian matrices.

Theorem 6.1. *Let $\{u_i\} \in W_h$ be the sequence solution to (5.1–5.3) and let u be the function solution to (1.5–1.7). Then, under the assumptions of Theorem 5.2, there is a constant c independent of h and τ and such that*

$$\max_{0 \leq i \leq M} \|u_i - u(t_i)\|_{-1} \leq c\tau^\nu \tag{6.1}$$

with $\nu = \frac{1}{4}$ in any case and with $\nu = \frac{1}{2}$ if $\beta = 0$ or if property 6.1 is valid.

Proof. We use the notation

$$U_i = u_i - u(t_i), \quad i = 0, 1 \dots M, \tag{6.2}$$

for the error at step i and we set

$$\Phi(s) = \phi_i - \phi(s), \quad s \in (t_{i-1}, t_i], \quad i = 1, 2 \dots M. \tag{6.3}$$

Combining (1.5) with (5.2) and using the definition (3.26) of B , we get a relationship between U_i and U_{i-1} for any $i = 1, 2 \dots M$. It reads

$$(\xi, BU_i)_1 = (\xi, BU_{i-1})_1 + \int_{t_{i-1}}^{t_i} ((\Phi(s), \xi)_1 + (g(u_{i-1}) - g(u(s)), \xi)_{0,h}) ds, \quad \xi \in V \tag{6.4}$$

and it is clear because of (1.7) and (5.1) that

$$U_0 = 0. \tag{6.5}$$

We take $\xi = BU_i$ as a test function in (6.4) and we bound the first term in the right-hand side thanks to Cauchy-Schwarz and Young's inequalities. We also use that $(\Phi(s), BU_i)_1 = (U_i, \Phi(s))_{0,h}$ (see (3.26)). We come to the conclusion that

$$|BU_i|_1^2 \leq |BU_{i-1}|_1^2 + R_i, \quad i = 1, 2 \dots M, \tag{6.6}$$

where

$$R_i = \int_{t_{i-1}}^{t_i} 2((U_i, \Phi(s))_{0,h} + (g(u_{i-1}) - g(u(s)), BU_i)_{0,h}) ds, \quad i = 1, 2 \dots M. \tag{6.7}$$

We will achieve the proof by using the stability results stated in Theorem 4.1 (for the o.d.e.) and in Theorem 5.2 (for the scheme). Our idea is to prove that there are two constants C_1, C_2 , independent of h and τ , as well as non-negative numbers r_i such that it holds

$$R_i \leq C_1 \tau |BU_{i-1}|_1^2 + r_i, \quad i = 1, 2 \dots M \tag{6.8}$$

as well as

$$\sum_{i=1}^{i=M} r_i \leq C_2 \tau^{2\nu}. \tag{6.9}$$

with $\nu = \frac{1}{4}$ and with $\nu = \frac{1}{2}$ if $\beta = 0$ or if property 6.1 holds. In view of (6.8), (6.9) and of the initial condition (6.5), the Gronwall Lemma applied to (6.6) actually establishes that

$$\max_{1 \leq i \leq M} |BU_i|_1^2 \leq (C_2 e^{C_1 T}) \tau^{2\nu} \tag{6.10}$$

and (6.1) follows. We just have to observe that the left-hand side of (6.10) is an estimate for $b_1^2 \max_{1 \leq i \leq M} |U_i|_{-1}^2$ as a consequence of the property (3.30) of B .

To prove (6.8–6.9), we first decompose the first term of the integrand in the definition (6.7) for R_i as the sum of three terms,

$$(U_i, \Phi(s))_{0,h} = (u_i - u_{i-1}, \Phi(s))_{0,h} + (u(s) - u(t_i), \Phi(s))_{0,h} + (u_{i-1} - u(s), \Phi(s))_{0,h} \tag{6.11}$$

and we bound them separately. At first, it follows from the definition (2.1) of the norm $\|\cdot\|_{-1,h}$ and from the Young's inequality that

$$(u_i - u_{i-1}, \Phi(s))_{0,h} \leq \frac{1}{2\tau} \|u_i - u_{i-1}\|_{-1,h}^2 + \frac{\tau}{2} \|\Phi(s)\|_1^2, \quad s \in (t_{i-1}, t_i] \tag{6.12}$$

and that

$$(u(s) - u(t_i), \Phi(s))_{0,h} \leq \frac{1}{2\tau} \|u(s) - u(t_i)\|_{-1,h}^2 + \frac{\tau}{2} \|\Phi(s)\|_1^2, \quad s \in (t_{i-1}, t_i].$$

Because of the Cauchy-Schwarz inequality and since u is differentiable, the first term in the right-hand side of the above inequality is not larger than $\frac{1}{2} \|\dot{u}\|_{L^2(t_{i-1}, t_i; H_h^{-1}(\Omega))}^2$, hence

$$(u(s) - u(t_i), \Phi(s))_{0,h} \leq \frac{1}{2} \|\dot{u}\|_{L^2(t_{i-1}, t_i; H_h^{-1}(\Omega))}^2 + \frac{\tau}{2} \|\Phi(s)\|_1^2, \quad s \in (t_{i-1}, t_i]. \tag{6.13}$$

Finally and because of equation (1.6) for $\phi(s)$, (5.2) for ϕ_i and (6.3) for $\Phi(s)$, we have, for any $s \in (t_{i-1}, t_i]$,

$$(u_{i-1} - u(s), \Phi(s))_{0,h} = (\partial S(u_{i-1}) - \partial S(u(s)), u_{i-1} - u(s))_0 - \beta(u_i - u_{i-1}, u_{i-1} - u(s))_0. \tag{6.14}$$

We use property (3.7) to bound the first term in the right-hand side and we make use of a proper Young's inequality to estimate the second one. We get that

$$\begin{aligned} (u_{i-1} - u(s), \Phi(s))_{0,h} &\leq -\frac{1}{\omega} \|\partial S(u_{i-1}) - \partial S(u(s))\|_0^2 \\ &\quad + \frac{\beta\tau^{\frac{1}{2}}}{2} \|u_{i-1} - u(s)\|_0^2 + \frac{\beta}{2\tau^{\frac{1}{2}}} \|u_i - u_{i-1}\|_0^2, \quad s \in (t_{i-1}, t_i]. \end{aligned} \tag{6.15}$$

However, if the non-degenerate entropy property 6.1 holds, it follows from (6.14) that

$$\begin{aligned} (u_{i-1} - u(s), \Phi(s))_{0,h} &\leq -\frac{1}{2\omega} \|\partial S(u_{i-1}) - \partial S(u(s))\|_0^2 \\ &\quad - \frac{\nu}{2} \|u_{i-1} - u(s)\|_0^2 - \beta(u_i - u_{i-1}, u_{i-1} - u(s))_0 \end{aligned}$$

and we can obtain a better estimate than (6.15) for $(u_{i-1} - u(s), \Phi(s))_{0,h}$. Using a convenient Young's inequality to bound $(u_i - u_{i-1}, u_{i-1} - u(s))_0$ yields

$$(u_{i-1} - u(s), \Phi(s))_{0,h} \leq -\frac{1}{2\omega} \|\partial S(u_{i-1}) - \partial S(u(s))\|_0^2 + C \|u_i - u_{i-1}\|_0^2 \tag{6.16}$$

for some $C < \infty$ which only depends on β and on $\nu > 0$.

We now treat the second term of the integrand in the definition (6.7) for R_i . It holds

$$(g(u_{i-1}) - g(u(s)), BU_i)_{0,h} \leq \|g(u_{i-1}) - g(u(s))\|_{-1,h} |BU_i|_1.$$

Using (3.22) to estimate $\|g(u_{i-1}) - g(u(s))\|_{-1,h}$ by mean of $\|g(u_{i-1}) - g(u(s))\|_0$ and taking the Lipschitz property of g (see (3.11)) into account, we deduce from the Young's inequality that there is a constant $C < \infty$ independent of h and τ and such that

$$(g(u_{i-1}) - g(u(s)), BU_i)_{0,h} \leq \frac{1}{2\omega} \|\partial S(u_{i-1}) - \partial S(u(s))\|_0^2 + C |BU_i|_1^2. \tag{6.17}$$

Because of the definition (6.2) for $U_j, j = i, i - 1$, we may write that

$$U_i = U_{i-1} + (U_i - U_{i-1}) = U_{i-1} + (u_i - u_{i-1}) + (u(t_i) - u(t_{i-1})).$$

We then use the Minkowski's inequality as well as the relation (3.29) to express $|B(u_i - u_{i-1})|_1$ and $|B(u(t_i) - u(t_{i-1}))|_1$ as $\|u_i - u_{i-1}\|_{-1,h}$ and $\|u(t_i) - u(t_{i-1})\|_{-1,h}$. We conclude that the constant C independent of h and τ involved in (6.17) may be chosen large enough so that $(g(u_{i-1}) - g(u(s)), BU_i)_{0,h}$ is not larger than

$$\frac{1}{2\omega} \|\partial S(u_{i-1}) - \partial S(u(s))\|_0^2 + C|BU_{i-1}|_1^2 + C\|u_i - u_{i-1}\|_{-1,h}^2 + C\|u(t_i) - u(t_{i-1})\|_{-1,h}^2.$$

From the Cauchy-Schwarz inequality and since u is differentiable, we deduce that $\|u(t_i) - u(t_{i-1})\|_{-1,h}^2$ is smaller than $\tau \|\dot{u}\|_{L^2(t_{i-1}, t_i; H_h^{-1}(\Omega))}^2$ and we conclude that

$$(g(u_{i-1}) - g(u(s)), BU_i)_{0,h} \leq \frac{1}{2\omega} \|\partial S(u_{i-1}) - \partial S(u(s))\|_0^2 + C \left(|BU_{i-1}|_1^2 + \|u_i - u_{i-1}\|_{-1,h}^2 + \tau \|\dot{u}\|_{L^2(t_{i-1}, t_i; H_h^{-1}(\Omega))}^2 \right). \tag{6.18}$$

We now add the three inequalities (6.18), (6.12), (6.13), and (6.15) or (6.16) if property 6.1 holds. We also take (6.11) into account and we use (3.22) to bound $\|u_i - u_{i-1}\|_{-1,h}$ by $\|u_i - u_{i-1}\|_0$. We conclude that the integrand $2((U_i, \Phi(s))_{0,h} + g(u_{i-1}) - g(u(s)), BU_i)_{0,h}$ in the definition (6.7) of R_i is less than

$$C|BU_{i-1}|_1^2 + C\|\dot{u}\|_{L^2(t_{i-1}, t_i; H_h^{-1}(\Omega))}^2 + C\tau \|\Phi(s)\|_1^2 + C\|u_i - u_{i-1}\|_0^2 + \beta\tau^{\frac{1}{2}} \|u_{i-1} - u(s)\|_0^2 + \frac{\beta}{\tau^{\frac{1}{2}}} \|u_i - u_{i-1}\|_0^2,$$

where $C < \infty$ does not depend on h and τ and where the last two terms may be omitted if property 6.1 is fulfilled. We use the Minkowski's inequality to estimate $\|\Phi(s)\|_1 = |\phi_i - \phi(s)|_1$ by $|\phi_i|_1 + |\phi(s)|_1$ and $\|u_{i-1} - u(s)\|_0$ by $\|u_{i-1}\|_0 + \|u(s)\|_0$. It follows that $R_i, i = 1, 2 \dots M$ satisfies (6.8) with $C_1 = C$ and with

$$r_i = C_0\tau \|\dot{u}\|_{L^2(t_{i-1}, t_i; H_h^{-1}(\Omega))}^2 + C_0\tau^2 |\phi_i|_1^2 + C_0\tau \|\phi\|_{L^2(t_{i-1}, t_i; H_0^1(\Omega))}^2 + C_0\tau \|u_i - u_{i-1}\|_0^2 + \beta\tau^{\frac{1}{2}} (\tau \|u_{i-1}\|_0^2 + \|u\|_{L^2(t_{i-1}, t_i; L^2(\Omega))}^2) + \|u_i - u_{i-1}\|_0^2 \tag{6.19}$$

for some constant C_0 independent of h and τ . If property 6.1 holds, the last term between brackets may be omitted in the right-hand side of (6.19).

To achieve the proof of the Theorem, it remains to find a constant C_2 independent of h and τ such that the r_i fulfil (6.9) with $\nu = \frac{1}{4}$ in general and with $\nu = \frac{1}{2}$ if $\beta = 0$ or if property 6.1 is valid. Since it follows from (6.19) that

$$\sum_{i=1}^{i=M} r_i \leq C_0\tau \left(\|\dot{u}\|_{L^2(0,T; H_h^{-1}(\Omega))}^2 + \tau \sum_{i=1}^{i=M} |\phi_i|_1^2 + \|\phi\|_{L^2(0,T; H_0^1(\Omega))}^2 + \sum_{i=1}^{i=M} \|u_i - u_{i-1}\|_0^2 \right) + \beta\tau^{\frac{1}{2}} \left(\sum_{i=1}^{i=M} \|u_i - u_{i-1}\|_0^2 + \tau \sum_{i=1}^{i=M} \|u_{i-1}\|_0^2 + \|u\|_{L^2(0,T; L^2(\Omega))}^2 \right), \tag{6.20}$$

and since the term of order $\tau^{\frac{1}{2}}$ may be left out under property 6.1, the conclusion can be directly derived from the stability estimate (4.1) for $\|\dot{u}\|_{L^2(0,T; H_h^{-1}(\Omega))}, \|\phi\|_{L^2(0,T; H_0^1(\Omega))}$ and $\|u\|_{L^2(0,T; L^2(\Omega))}$ and from the stability estimate (5.12) for $\sum_{i=1}^{i=M} \|u_i - u_{i-1}\|_0, \tau \sum_{i=1}^{i=M} |\phi_i|_1^2$ and $\tau \sum_{i=1}^{i=M} \|u_{i-1}\|_0^2$. □

Remark 6.1. When the order of convergence ν is only $\frac{1}{4}$, the error estimate (6.1) for the integration method (5.1–5.3) is *not optimal*. The reason is as follows.

Because of Theorem 4.1 and inequality (3.22), \dot{u} is bounded in $L^2(0, T; H^{-1}(\Omega))$. Hence it holds

$$\|u(t_2) - u(t_1)\|_{-1} \leq C\tau^{\frac{1}{2}}, \quad t_1, t_2 \in [0, T] \quad (6.21)$$

for some constant C , independent on h . The Hölder property (6.21) has a consequence for the best interpolant \hat{u} of u in the space of functions that are piecewise constant between the time steps $t_0 < t_1 < t_2 \dots$. It implies that \hat{u} converges to u at least like $\tau^{\frac{1}{2}}$ in $L^\infty(0, T; H^{-1}(\Omega))$, meaning that estimate (6.1) with $\nu = \frac{1}{4}$ is suboptimal.

REFERENCES

- [1] A. Friedman, The Stefan problem in several space variables. *Trans. Amer. Math. Soc.* **132** (1968) 51–87.
- [2] A.E. Berger, H. Brezis and J.C.W. Rogers, A numerical method for solving $u_t - \Delta f(u) = 0$. *RAIRO. Anal. Numér.* **13** (1979) 297–312.
- [3] C.M. Elliott, Error analysis of the enthalpy method for the Stefan problem. *IMA J. Numer. Anal.* **7** (1987) 61–71.
- [4] S.R. De Groot and P. Mazur, *Non-equilibrium thermodynamics*. North-Holland, Amsterdam (1962).
- [5] H. Brezis, *Analyse fonctionnelle, Théorie et applications*. Masson, Paris (1993).
- [6] H.W. Alt and S. Luckhaus, Quasilinear elliptic-parabolic differential equations. *Math. Z.* **183** (1983) 311–341.
- [7] I. Prigogine, *Thermodynamics of irreversible processes*. Interscience Publ. (1967).
- [8] J.D.P. Donnelly, A model for non-equilibrium thermodynamic processes involving phase changes. *J. Inst. Math. Appl.* **24** (1979) 425–438.
- [9] J.F. Ciavaldini, Analyse numérique d'un problème de Stefan à deux phases par une méthode d'éléments finis. *SIAM J. Numer. Anal.* **12** (1975) 464–487.
- [10] J.W. Jerome and M.E. Rose, Error estimates for the multidimensional two-phase Stefan problem. *Math. Comp.* **39** (1982) 377–414.
- [11] K. Yosida, *Functional Analysis*. Springer-Verlag, Berlin (1984).
- [12] E. Magenes, Remarques sur l'approximation des problèmes paraboliques non-linéaires, in *Analyse Mathématique et Applications*, Gauthier-Villars, Paris (1988) 297–318.
- [13] E. Magenes, R.H. Nochetto and C. Verdi, Energy error estimates for a linear scheme to approximate nonlinear parabolic problems. *RAIRO. Modél. Math. Anal. Numér.* **21** (1987) 655–678.
- [14] M. Crouzeix and A.L. Mignot, *Analyse numérique des équations différentielles*. Masson (1989).
- [15] G.H. Meyer, Multidimensional Stefan problems. *SIAM J. Numer. Anal.* **10** (1973) 522–538.
- [16] O. Krüger, Modélisation et analyse numérique de problèmes de réaction-diffusion provenant de la solidification d'alliages binaires. Technical Report 2071, Thèse EPFL (1999).
- [17] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*. Pitman, Boston (1985).
- [18] M. Paolini, G. Sacchi and C. Verdi, Finite element approximations of singular parabolic problems. *Internat. J. Numer. Methods Engrg.* **26** (1988) 1989–2007.
- [19] P.G. Ciarlet, *The Finite Element Method for Elliptic Problem*. North Holland, Amsterdam (1978).
- [20] J. Rulla, Error analysis for implicit approximations to Cauchy problems. *SIAM J. Numer. Anal.* **33** (1996) 68–87.
- [21] V. Thomée, *Galerkin finite element methods for Parabolic Problems*. Springer-Verlag, Berlin (1984).
- [22] W. Jäger and J. Kačur, Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes. *RAIRO. Modél. Math. Anal. Numér.* **29** (1995) 605–627.