Weak solutions to Friedrichs systems with convex constraints

Bruno Després*, Frédéric Lagoutière† and Nicolas Seguin‡

December 31, 2010

MSC Classification: 35L40, 35L45, 35L65, 35L67, 35L85.

Abstract

We are interested in a problem arising for instance in elastoplasticity modeling, which consists in a system of partial differential equations *and* a constraint specifying that the solution should remain, for every time and every position, in a certain set. This constraint is generally incompatible with the invariant domains of the original model, thus this problem has to be precised in mathematical words. We here follow the approach proposed in [8] that furnishes a weak formulation of the constrained problem *à la* Kruzhkov. More precisely, the present paper deals with the study of the well-posedness of Friedrichs systems under convex constraints, in any space dimension. We prove that there exists a unique weak solution, continuous in time, square integrable in space, and with values in the constraints domain. This is done with the use of a discrete approximation scheme: we define a numerical approximate solution and prove, thanks to compactness properties, that it converges toward a solution to the constrained problem. Uniqueness is proven *via* energy (or entropy) estimates. Some numerical illustrations are provided.

1 Introduction

The understanding of constraints in Mathematical Physics has always known a strong impetus and the development of an adapted framework for non-linear partial differential equations is still a great challenge. In the context of the mathematical theory of plasticity one may think for example about the seminal works [10, 27] and about all the works inspired by these references, see also [7]:

*Université Pierre et Marie Curie-Paris6, UMR 7598, LJLL, Paris, F-75005 France. despres@ann.jussieu.fr

†Université Paris-Sud 11, Département de Mathématiques, CNRS UMR 8628, Bâtiment 425, 91405 Orsay Cedex, France, and Équipe-Projet SIMPAF, Centre de Recherche INRIA Futurs, Parc Scientifique de la Haute Borne, 40, Avenue Halley B.P. 70478, F-59658 Villeneuve d'Ascq Cedex, France. frederic.lagoutiere@math.u-psud.fr

‡Université Pierre et Marie Curie-Paris6, UMR 7598, LJLL, Paris, F-75005 France. nicolas.seguin@upmc.fr

to our knowledge, this approach is variational by nature and restricted to smooth solutions or at least to solutions which are differentiable with respect to the time variable. Motivated by applications in compressible plasticity with shocks, thus leading to discontinuous solutions (one may find references in [15, 3, 8, 23]), we aim at developing a mathematical framework for constrained weak solutions of hyperbolic equations [17, 25]. When the constraint is saturated the reduced equation is no-linear and non-conservative in most cases. It means that the definition of weak solutions may be incompatible with the non-uniqueness of solutions which is a characteristic of shock solutions for non-conservative formulations [18, 19, 20, 1]. It is possible to add hypotheses of maximal dissipation to make a selection between all possible generalized solutions, see [15, 24, 6] and references therein: however one may consider this approach as somewhat artificial in our context.

In this work we study Friedrichs systems which are linear equations, endowed with a non-linear constraint. Our main result, theorem 3 at the end of this preliminary section, states that Friedrichs systems with a general convex constraint admit weak solutions, and that these solutions are unique. The key of the proof is a new weak formulation of the problem. The discussion of discontinuous weak solutions can be made on the basis of the weak Rankine Hugoniot relation, which is an algebraic relation deduced from the weak formulation, see theorem 20. Roughly speaking, if K, the convex set of admissible states, is non-empty, then the non-conservative equations (that one can write for constrained solutions on the boundary of K) cannot develop "exotic" non-conservative shocks. It is therefore neither necessary to rely on the theory of non-conservative products [18, 19, 20] nor on a maximal dissipation principle to study weak solutions for constrained Friedrichs systems.

We consider the Cauchy problem

$$\left\{\begin{array}{ll} \partial_t u + \displaystyle\sum_{i=0}^{d} A_i \partial_{x_i} u = 0, & (t,x) \in \mathbb{R}_+ \times \mathbb{R}^d, \\ u(t,x) \in K, & (t,x) \in \mathbb{R}_+ \times \mathbb{R}^d, \\ u(0,x) = u^0(x), & x \in \mathbb{R}^d, \end{array}\right. \tag{1}$$

where K is a non-empty closed convex subset of $\mathbb{R}^n$, the matrices $A_i \in \mathcal{M}_n(\mathbb{R})$ are supposed to be symmetric and the unknown is $u \in \mathbb{R}^n$ $(n, d \in \mathbb{N}^*)$. Of course the equation above might be incompatible with the request $u \in K$, thus this problem has to be modeled first.

It is worth noting that the Cauchy problem is invariant with respect to translations in the space state: consider a solution u of the Cauchy problem (1) associated with a convex K and with the initial data u^0; then for all constant vector $U \in \mathbb{R}^n$, $u - U$ is a solution of the Cauchy problem (1) associated with the convex $K - U$ and with the initial data $u^0 - U$. Therefore, we assume all along this paper that

$$0 \in K$$

without restriction on the results which follow.

In order to approximate the problem (1), it is natural to consider the relaxation system (more details are provided in section 2)

$$\begin{cases} \partial_t u_\varepsilon + \sum_{i=0}^{d} A_i \partial_{x_i} u_\varepsilon = \dfrac{1}{\varepsilon} \left(P_K \left(u_\varepsilon \right) - u_\varepsilon \right), & (t,x) \in \mathbb{R}_+ \times \mathbb{R}^d, \\ u_\varepsilon(0,x) = u^0(x), & x \in \mathbb{R}^d, \end{cases} \tag{2}$$

where the operator $P_K : \mathbb{R}^n \to K$ is the projection onto K, that is to say

$$\langle u - v, P_K(u) - v \rangle \geq 0 \quad \forall v \in K, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product of $\mathbb{R}^n$. The initial data is assumed to be in K,

$$u^0(x) \in K \text{ for almost every } x \in \mathbb{R}^d.$$

If there exists u such that $u_\varepsilon \to u$ almost everywhere when $\varepsilon \to 0$, one may expect that u belongs to K since $P_K u = u$. Our main objective is to characterize the limit u (see the appendix for more details on the justification of this limit). Problem (2) can be seen as a simplification of a more general problem arising in many situations, which consists in the characterization of weak solutions of hyperbolic systems of conservation laws with constraints. Since the limit problem may be non-linear, a notion of admissible solution must be given. For this purpose, we add viscous terms in (2) and obtain (noting $\Delta = \sum_{i=1}^{d} \partial_{x_i}^2$)

$$\partial_t u_{\varepsilon,\nu} + \sum_{i=1}^{d} A_i \partial_{x_i} u_{\varepsilon,\nu} = \frac{1}{\varepsilon} \left(P_K \left(u_{\varepsilon,\nu} \right) - u_{\varepsilon,\nu} \right) + \nu \Delta u_{\varepsilon,\nu}. \tag{4}$$

We will introduce a specific weak formulation of Problem (4). Let $\kappa \in K$ be a test vector (that is κ is a constant vector independent of the time variable t and of the space variable x). Consider a smooth solution of (4) and take the scalar product with $u_{\varepsilon,\nu} - \kappa$. One gets

$$\begin{aligned} &\partial_t \frac{|u_{\varepsilon,\nu}|^2}{2} + \sum_{i=1}^{d} \partial_{x_i} \frac{\langle u_{\varepsilon,\nu}, A_i u_{\varepsilon,\nu} \rangle}{2} - \partial_t \langle \kappa, u_{\varepsilon,\nu} \rangle - \partial_{x_i} \langle \kappa, A_i u_{\varepsilon,\nu} \rangle \\ &= \frac{1}{\varepsilon} \langle u_{\varepsilon,\nu} - \kappa, P_K \left(u_{\varepsilon,\nu} \right) - u_{\varepsilon,\nu} \rangle - \nu \sum_{i=1}^{d} |\partial_{x_i} u_{\varepsilon,\nu}|^2 + \nu \Delta \frac{|u_{\varepsilon,\nu}|^2}{2} - \nu \Delta \langle \kappa, u_{\varepsilon,\nu} \rangle \\ &\leq \nu \Delta \frac{|u_{\varepsilon,\nu}|^2}{2} - \nu \Delta \langle \kappa, u_{\varepsilon,\nu} \rangle . \end{aligned}$$

Let $T > 0$, define the space of non-negative smooth test functions with compact support

$$\mathcal{D}_T^+ = \{ \varphi \in \mathcal{C}_0^\infty \left([0, T[\times \mathbb{R}^n \right), \varphi(t,x) \geq 0 \text{ for all } (t,x) \in [0, T[\times \mathbb{R}^n \} .$$

Let us emphasize that φ vanishes for $t = T$ and has compact support in space, but $x \mapsto \varphi(0, x)$ does not necessarily vanish. Let us multiply the previous inequality by a test function $\varphi \in \mathcal{D}_T^+$ and integrate over $[0, T] \times \mathbb{R}^d$. We get

$$\int_0^T \int_{\mathbb{R}^d} \left[\left(\frac{|u_{\varepsilon,\nu}|^2}{2} - \langle \kappa, u_{\varepsilon,\nu} \rangle \right) \partial_t + \sum_{i=1}^d \left(\frac{\langle u_{\varepsilon,\nu}, A_i u_{\varepsilon,\nu} \rangle}{2} - \langle \kappa, A_i u_{\varepsilon,\nu} \rangle \right) \partial_{x_i} \right] \varphi \, dx \, dt$$
$$+ \int_{\mathbb{R}^d} \left(\frac{|u^0|^2}{2} - \langle \kappa, u^0 \rangle \right) \varphi(0, x) \, dx \geq \nu \int_0^T \int_{\mathbb{R}^d} \left(\frac{|u_{\varepsilon,\nu}|^2}{2} - \langle \kappa, u_{\varepsilon,\nu} \rangle \right) \Delta \varphi \, dx \, dt.$$

Notation 1. *Throughout this work we will use for the sake of simplicity the notation L^2 instead of $(L^2)^n$, n being the dimension of the unknown vector u. We will also use H^1 instead of $(H^1)^n$.*

Assume that $u_{\varepsilon,\nu} \to u$ in L^2 when $\varepsilon \to 0^+$ and $\nu \to 0^+$, then the limit solution u satisfies the weak formulation

$$\int_0^T \int_{\mathbb{R}^d} \left[\left(\frac{|u|^2}{2} - \langle \kappa, u \rangle \right) \partial_t \varphi + \sum_{i=1}^d \left(\frac{\langle u, A_i u \rangle}{2} - \langle \kappa, A_i u \rangle \right) \partial_{x_i} \varphi \right] dx \, dt$$
$$+ \int_{\mathbb{R}^d} \left(\frac{|u^0|^2}{2} - \langle \kappa, u^0 \rangle \right) \varphi^0 dx \geq 0, \quad (5)$$

for all $\kappa \in K$ and $\varphi \in \mathcal{D}_T^+$. For technical reasons, let us add vanishing terms which are derivatives of κ, multiply by a factor 2 and consider the following (equivalent) characterization:

Definition 2. *Let $u^0 \in L^2(\mathbb{R}^d, K)$. A function u is a* weak constrained solution *of (1) if $u \in L^2([0, T] \times \mathbb{R}^d, K)$ and satisfies for every $\kappa \in K$ and $\varphi \in \mathcal{D}_T^+$*

$$\int_0^T \int_{\mathbb{R}^d} \left[|u - \kappa|^2 \partial_t \varphi + \sum_{i=1}^d \langle u - \kappa, A_i (u - \kappa) \rangle \, \partial_{x_i} \varphi \right] dt \, dx$$
$$+ \int_{\mathbb{R}^d} |u^0 - \kappa|^2 \varphi(0, x) \, dx \geq 0. \quad (6)$$

The main result of the paper is

Theorem 3. *Assume that $u^0 \in L^2(\mathbb{R}^d, K)$. There exists a unique weak constrained solution $u \in L^2([0, T] \times \mathbb{R}^d, K)$ to (1) in the sense of Definition 2. Besides, this solution also belongs to $\mathcal{C}([0, T], L^2(\mathbb{R}^d, K))$ and if we assume that $u^0 \in H^1(\mathbb{R}^d, K)$, then $u \in L^\infty([0, T], H^1(\mathbb{R}^d, K))$.*

The proof of theorem 3 is based on a generalization of the Kruzhkov entropy technique and the construction of a discrete solution. The uniqueness of weak solutions in $L^2([0, T] \times \mathbb{R}^d, K)$ is a consequence of lemma 9. The regularity for $u_0 \in H^1(\mathbb{R}^d, K)$ is a consequence of lemma 10. The main part of the theorem, that is the existence in $\mathcal{C}([0, T], L^2(\mathbb{R}^d, K))$, is proven in section 4.

This work is organized as follows. In a first section we discuss various examples, some of them arising in material strength modeling. Next, we focus on Friedrichs systems and Problem (1). Uniqueness is proven, by the use of the Kruzhkov's method of doubling of variables [14]. In section 4, a numerical scheme is constructed and passing the to the limit for the associated numerical solution, we prove the existence of at least one weak constrained solution. Then, discontinuous solutions are studied in details, see theorem 20. At the end, we propose two numerical examples. The first one concerns the wave equation with a simple set of constraint while the second one corresponds to the physical situation where shocks propagate in a bar with isotropic strain hardening. In Appendix, the convergence of the relaxation approximation is tackled.

Remark 4. *Several improvements of theorem 3 and of the results which follow should be given but, for the sake of clarity, we have preferred to avoid as much as possible any technical issue. As an example, one could have dropped the assumptions* $u^0 \in K$ *and* $0 \in K$.

Concerning the numerical scheme, we have chosen to consider the classical Rusanov scheme on Cartesian meshes with a splitting technique for the constraint. The case of unstructured meshes with a general class of Finite Volume methods is investigated in [29, 13]. However, the mathematical tools are rather different (and more complicated), without providing any interesting feature in our context. That is why we restrict our analysis to the Rusanov scheme on Cartesian meshes. At last, let us remark that the study of the convergence of the solutions of the relaxation approximation (2) *to the weak constrained solutions is interesting by itself. It is partially discussed in the appendix.*

Remark 5. *The case of a scalar conservation law, i.e.* $n = 1$*, is trivial since a convex set in* $\mathbb{R}$ *is an interval and the entropy solutions of scalar conservation laws satisfy a maximum principle that guaranties that it remains in the interval.*

2 Examples

Many basic or more physical examples fit in the framework studied in this work. We arbitrarily split these examples in two categories. The first category concerns linear wave systems endowed with various constraints. The second category deals with problems that come from non-linear physics.

2.1 Wave system plus constraint

Let us consider the system in one space dimension

$$\left\{ \begin{array}{l} \partial_t u + \partial_x v = 0, \\ \partial_t v + \partial_x u = 0. \end{array} \right. \tag{7}$$

In what follows we associate (7) with different types of constraints.

2.1.1 A ball-shaped constraint space

We endow (7) with the constraint $(u, v) \in K_1$ with

$$K_1 = \{(\widetilde{u}, \widetilde{v}),\ \widetilde{u}^2 + \widetilde{v}^2 \leq 1\}. \tag{8}$$

This constraint is very close to a Von Mises plasticity constraint [9] if one thinks about material strength modeling (that is elastic and plastic modeling in our case) It means that (u, v) are two components of the stress tensor. A possibility to incorporate this constraint in the system is to consider

$$\left\{ \begin{array}{l} \partial_t u_\varepsilon + \partial_x v_\varepsilon = -\frac{1}{\varepsilon} \left(u^\star(u_\varepsilon, v_\varepsilon) - u_\varepsilon\right), \\ \partial_t v_\varepsilon + \partial_x u_\varepsilon = -\frac{1}{\varepsilon} \left(v^\star(u_\varepsilon, v_\varepsilon) - v_\varepsilon\right), \end{array} \right. \tag{9}$$

where

$$u^\star(a, b) = \frac{a}{\max\left(1, a^2 + b^2\right)} \text{ and } v^\star(a, b) = \frac{b}{\max\left(1, a^2 + b^2\right)}.$$

The vector $(u^\star(a, b), v^\star(a, b))$ is the Euclidean projection onto K_1.

2.1.2 A square-shaped constraint space

We consider the same problem but the convex is now the unit square

$$K_2 = \{(\widetilde{u}, \widetilde{v}),\ \max(|\widetilde{u}|, |\widetilde{v}|) \leq 1\}. \tag{10}$$

If we continue to make a parallel with material strength modeling, K_2 is a normalized Tresca [9] constraint.

2.1.3 A linear constraint space

It is possible to consider many different types of convex constraints, still for the same initial wave system: we single out the trivial case

$$K_3 = \{(\widetilde{u}, \widetilde{v}),\ \widetilde{u} = \alpha \widetilde{v}\}, \qquad \alpha \in \mathbb{R}.$$

Another example in the same family is the system

$$\left\{ \begin{array}{l} \partial_t u + \partial_x v = 0, \\ \partial_t v + \partial_x u - \partial_x w = 0, \\ \partial_t w - \partial_x v = 0 \end{array} \right. \tag{11}$$

endowed with the linear constraint $w - 2u = 0$. This example comes from [5] and has been designed to explain that a crude introduction of the constraint $w - 2u = 0$ may result in a ill-posed system. Indeed if one eliminates w directly in the second equation, one obtains

$$\left\{ \begin{array}{l} \partial_t u + \partial_x v = 0, \\ \partial_t v - \partial_x u = 0 \end{array} \right. \tag{12}$$

which is ill-posed.

On the basis of the weak formulation (6) it is immediate to recover the approach proposed in [5, 2, 21]. The suitable general hypothesis is that K is a linear subspace, that is

$$K = \mathrm{Span}\,(Z_j)_{1\leq j<n}\,,\ Z_j \in \mathbb{R}^n, \quad \dim(K) = p < n. \tag{13}$$

Lemma 6. *Assume that K is a linear subspace defined by (13) and that the basis is orthonormal: $\langle Z_i, Z_j\rangle = \delta_{ij}$. Then the solutions of the weak formulation (6) are weak solutions of the linear subsystem*

$$\partial_t \widetilde{u} + \sum_{i=1}^{d} \widetilde{\mathbf{A}_i} \partial_{x_i} \widetilde{u} = 0, \tag{14}$$

where $\widetilde{u} = \big(\langle Z_j, u\rangle\big)_{1\leq j\leq p} \in \mathbb{R}^p$ and $\widetilde{\mathbf{A}_i} = \big(\langle Z_j, A_i Z_{j'}\rangle\big)_{1\leq j,j'\leq p} = \widetilde{\mathbf{A}_i}^{\,t} \in \mathbb{R}^{p\times p}$.

Proof. Let us first recall that the weak formulation (6) is equivalent to (5). For the simplicity of the proof we consider a test function φ which vanishes at time $t = 0$. Consider a test vector $\kappa = \lambda Z$ where $Z \in K$ and λ is a real number. So

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}_+} \left[\left(\frac{|u|^2}{2} - \lambda\,\langle Z, u\rangle\right)\partial_t\varphi + \sum_{i=1}^{d}\left(\frac{\langle u, A_i u\rangle}{2} - \lambda\,\langle Z, A_i u\rangle\right)\partial_{x_i}\varphi\right] dt\ dx \geq 0$$

for all $Z \in K$. Taking $\lambda = 0$, one gets the inequality

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}_+} \left[\frac{|u|^2}{2}\partial_t\varphi + \sum_{i=1}^{d}\frac{\langle u, A_i u\rangle}{2}\partial_{x_i}\varphi\right] dt\ dx \geq 0, \tag{15}$$

and letting λ go to $\pm\infty$ one obtains the series of equalities

$$\left\langle Z, \int_{\mathbb{R}^d}\int_{\mathbb{R}_+} [u\ \partial_t\varphi + A_i u \partial_{x_i}\varphi]\, dt\ dx \right\rangle = 0 \tag{16}$$

for all admissible test vector Z and in particular for all Z_j.

Let us analyze the series of equalities (16). Set $\alpha_j = \langle Z_j, u\rangle$. Since the basis is orthonormal one can write that $u = \sum_{j'=1}^{p} \alpha_{j'} Z_{j'}$ where the coefficients $\alpha_{j'}$ are functions of the space and time coordinates. Therefore one has

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}_+} \left[\alpha_j\partial_t\varphi + \sum_{j'=1}^{p}\langle Z_j, A_i Z_{j'}\rangle\,\alpha_{j'}\partial_{x_i}\varphi\right] dt\ dx = 0,$$

which means that

$$\partial_t\alpha_j + \sum_{j'=1}^{p}\langle Z_j, A_i Z_{j'}\rangle\,\partial_{x_i}\alpha_{j'} = 0 \tag{17}$$

holds in the weak sense for all $1 \leq j \leq p$. It proves (14). □

Note that Inequality (15) is in fact an equality. Indeed, multiplying (17) by α_j and summing over all $1 \leq j \leq p$, one gets

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}_+} \left[\frac{|u|^2}{2} \partial_t \varphi + \sum_{i=1}^{d} \frac{\langle u, A_i u \rangle}{2} \partial_{x_i} \varphi \right] dt \; dx = 0.$$

The reason why (16) implies (15) is the following well known property: the entropy inequality is an equality for linear systems.

2.2 Models for material strength

The mathematical theory for material strength contains many of open mathematical problems [15, 24]. Some of them can be addressed with the tools developed in this work.

2.2.1 Perfect plasticity

Let us consider the following model problem [8] that intends to be representative of elastoplasticity models written in Eulerian coordinates in one dimension for planar flows (the velocity is parallel to the x direction) without shear (the unknowns depend only on x)

$$\left\{ \begin{array}{l} \partial_t \rho + \partial_x(\rho u) = 0, \\ \partial_t \rho u + \partial_x(\rho u^2 + p - s_1) = 0, \qquad p = p(\tau, \varepsilon), \; \tau = \frac{1}{\rho}, \\ \partial_t \rho s_1 + \partial_x(\rho u s_1 - \frac{4}{3}\alpha u) = 0, \\ \partial_t \rho s_2 + \partial_x(\rho u s_2 + \frac{2}{3}\alpha u) = 0, \\ \partial_t \rho s_3 + \partial_x(\rho u s_3 + \frac{2}{3}\alpha u) = 0, \\ \partial_t \rho \sigma_{23} + \partial_x(\rho u \sigma_{23}) = 0, \\ \partial_t \rho e + \partial_x(\rho u e + p u - s_1 u) = 0. \end{array} \right. \tag{18}$$

The unknowns are: ρ the density of mass, u the velocity, $s = (s_1, s_2, s_3, \sigma_{23})$ the deviatoric part of the stress tensor and e the total energy. The total energy is the sum of the internal energy, the kinetic energy and the elastic energy $e = \varepsilon + \frac{1}{2}u^2 + \frac{1}{4\alpha}(s_1^2 + s_2^2 + s_3^2 + 2\sigma_{23}^2)$. This system is closed by a pressure law $p = p(\rho, \varepsilon)$ where ε is the internal energy. The model parameter α is related to the Lame coefficient. More details can be found in [8].

A plastic material is not able to handle too large deviatoric stresses. It is modeled by some convex inequality constraint, for example

$$s_1^2 + s_2^2 + s_3^2 + 2\sigma_{23}^2 \leq k^2 \tag{19}$$

where $k \geq 0$ is a given constant. This is called perfect plasticity [3, 22, 9]. The classical analysis of relaxation to the convex of plasticity for isotropic materials [15] is as follows. We consider that perfect plasticity (18-19) is the limit $\lambda \to 0^+$

of the Maxwell's viscoelastic model

$$\left\{\begin{array}{ll} \partial_t\rho + \partial_x(\rho u) = 0, & \\ \partial_t\rho u + \partial_x(\rho u^2 + p - s_1) = 0, & p = p(\tau,\varepsilon), \\ \partial_t\rho s_1 + \partial_x(\rho u s_1 - \frac{4}{3}\alpha u) = -\frac{\mathbf{I}(U)}{\lambda}s_1, & \\ \partial_t\rho s_2 + \partial_x(\rho u s_2 + \frac{2}{3}\alpha u) = -\frac{\mathbf{I}(U)}{\lambda}s_2, & \\ \partial_t\rho s_3 + \partial_x(\rho u s_3 + \frac{2}{3}\alpha u) = -\frac{\mathbf{I}(U)}{\lambda}s_3, & \\ \partial_t\rho\sigma_{23} + \partial_x(\rho u\sigma_{23}) = -\frac{\mathbf{I}(U)}{\lambda}\sigma_{23}, & \\ \partial_t\rho e + \partial_x(\rho u e + pu - s_1 u) = -\mathbf{I}(U)\frac{s_1^2+s_2^2+s_3^2+2\sigma_{23}^2}{2\alpha\lambda}. & \end{array}\right. \tag{20}$$

The function $U \mapsto \mathbf{I}(U)$ is zero inside the domain of plasticity and is equal to one outside

$$\left\{\begin{array}{lll} \mathbf{I}(U) = 1 & \text{if} & s_1^2 + s_2^2 + s_3^2 + 2\sigma_{23}^2 > k^2, \\ \mathbf{I}(U) = 0 & \text{if} & s_1^2 + s_2^2 + s_3^2 + 2\sigma_{23}^2 < k^2. \end{array}\right.$$

At the limit $\lambda \to 0^+$ one recovers (formally) the constraint (19).

Let us now simplify the relaxation model (20) by retaining only the unknowns u, s_1, s_2, s_3 and $\widetilde{\sigma_{23}} = \sqrt{2}\sigma_{23}$. We also neglect the variations of the density and the effect of the transport. One gets the simplified model

$$\left\{\begin{array}{l} \partial_t u - \alpha\partial_x s_1 = 0, \\ \partial_t s_1 - \frac{4}{3}\alpha\partial_x u = -\frac{s_1^\star - s_1}{\varepsilon}, \\ \partial_t s_2 + \frac{2}{3}\alpha\partial_x u = -\frac{s_2^\star - s_2}{\varepsilon}, \\ \partial_t s_3 + \frac{2}{3}\alpha\partial_x u = -\frac{s_3^\star - s_3}{\varepsilon}, \\ \partial_t\widetilde{\sigma_{23}} = -\frac{\widetilde{\sigma_{23}}^\star - \widetilde{\sigma_{23}}}{\varepsilon}. \end{array}\right. \tag{21}$$

By inspection of the Von Mises constraint (19) one may check that the convenient definition of $(s_1^\star, s_2^\star, s_3^\star, \widetilde{\sigma_{23}}^\star)$ is

$$(s_1^\star, s_2^\star, s_3^\star, \widetilde{\sigma_{23}}^\star) = \Pi_4(s_1, s_2, s_3, \widetilde{\sigma_{23}})$$

where Π_4 is the projection onto the closed ball with radius k in $\mathbb{R}^4$. This problem is particular case of the general problem treated in this work.

2.2.2 Isotropic strain-hardening

This example is taken from [23]. We consider an elastic bar in one-dimensional configuration. The normalized unconstrained problem writes

$$\left\{\begin{array}{l} \partial_t u - \partial_x\sigma = 0, \\ \partial_t\sigma - \partial_x u = 0, \\ \partial_t\gamma = 0, \end{array}\right. \tag{22}$$

where the velocity in the bar is u, the uniaxial stress is σ and γ models the hardening. It is assumed that the bar is elastic if $|\sigma| + g(\gamma) < k$ where $k \geq 0$ is a given constant. The bar is in a plastic regime if $|\sigma| + g(\gamma) = k$. A convenient assumption is that function g is convex and

$$0 < \alpha \leq g'(\gamma) \leq \beta, \quad \text{with } g(0) = 0.$$

In Reference [23] the authors consider an initial condition $\gamma(t=0,x)=0$ and the modifications of the equation for γ is such that $\partial_t\gamma \leq 0$. Therefore only the branch $\gamma \leq 0$ matters in the definition of g. By construction $g(-\infty)=-\infty$, which enables us to rewrite the elasticity domain as

$$|\sigma| \leq k(\gamma) \equiv k - g(\gamma).$$

Physically, the domain of elasticity increases during the history of the material. This is called *isotropic hardening*. In [23] the authors endow the problem with the natural L^2 norm. Using our notations, the relaxation problem can be rewritten as

$$\partial_t U + A\partial_x U = -\frac{1}{\varepsilon}(P_K(U) - U) \tag{23}$$

with

$$U = (u, \sigma, \gamma), \quad A = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and

$$K = \left\{(u,\sigma,\gamma) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^-,\ \ |\sigma| + g(\gamma) \leq k\right\}. \tag{24}$$

The closed set K is obviously convex, so this problem can be dealt within our framework. Actually the result proven in [23] is a consequence of the main theorem of this work, with a H^1 initial data. Using the general results of the present work it is easy to discuss the solution of the elastic bar with strain hardening problem with discontinuous initial data, which is the case if for example the strain hardening is not constant initially (γ is discontinuous in space at $t=0$).

2.3 Saint-Venant plus mass loss

This non-linear system is proposed in [1]. It models water in a channel with flat bottom with the possibility of flooding outside the channel. It writes

$$\begin{cases} \partial_t h + \partial_x hu = Q, \\ \partial_t hu + \partial_x \left(hu^2 + \frac{g}{2}h^2\right) = Qu. \end{cases} \tag{25}$$

The height of water $h \geq 0$ is constrained to be less than 1. It can be done with the Lagrange multiplier Q. Indeed, considering a smooth solution, the constraint is satisfied by choosing Q in the following way:

- if $h < 1$ then $Q = 0$;
- if $h = 1$ then $Q = \partial_x hu = \partial_x u$.

This example is not covered by the theory developed in the present work because the system is non-linear. Nevertheless we quote [8] where a formalism is proposed in order to take into account such models. The idea is to use the adjoint or entropy variable in order to symmetrize the problem.

3 Uniqueness

It is important to notice that we are dealing with "symmetric" entropies and entropy flux (this is the interest of using (6) instead of (5)), so that the Kruzhkov's theory [14] naturally applies. Before studying the uniqueness of the constrained weak solution, let us briefly recall that such a solution satisfies the initial condition in the strong sense, even if weak solutions are not necessarily continuous, see definition 2. This result is needed for the proof of lemma 8.

Proposition 7. *Let u be a solution of the constrained problem in the sense of Definition 2. For all non-negative $\xi \in \mathcal{C}_0^\infty(\mathbb{R}^d)$, one has*

$$\operatorname*{ess\,lim}_{t\to 0} \int_{\mathbb{R}^d} |u(t,x) - u^0(x)|^2 \ \xi(x) \ dx = 0. \tag{26}$$

Proof. The proof of this classical result can be found in [4] and in references therein. It is based on a special version of the Kruzhkov's method of doubling of variables [14]. We only recall the main guidelines.

First, let us take a test function in $\mathcal{D}_T^+$ of the form $\varphi(t,x,y) = \xi(x)\eta_\alpha(t)\rho_\epsilon(x-y)$, $y \in \mathbb{R}^d$, where $\xi \in \mathcal{C}_0^\infty(\mathbb{R}^d)$ is a non-negative test function, $\rho_\epsilon(z)$ is a classical sequence of mollifiers and $\eta_\alpha(t) = \max(0, \min(1, (T-t+\alpha)/\alpha))$. Notice that $\eta'_\alpha(t) \leq 0$ and that $\eta'_\alpha \to -\delta_T$ as $\alpha \to 0^+$ (in the sense of distributions). Since $\partial_t\varphi$ is non-positive, one has

$$\begin{aligned}
&|u(t,x) - u^0(y)|^2 \partial_t\varphi(t,x,y) \leq |u(t,x) - u^0(x)|^2 \partial_t\varphi(t,x,y) \\
&+ |u^0(x) - u^0(y)|^2 \partial_t\varphi(t,x,y) - 2|u(t,x) - u^0(x)||u^0(x) - u^0(y)|\partial_t\varphi(t,x,y)
\end{aligned}$$

for almost every (t,x,y). As a consequence, integrating this inequality with respect to x, y and t provides

$$\begin{aligned}
&\int_0^{T+\alpha} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |u(t,x) - u^0(y)|^2 \partial_t\varphi(t,x,y) \ dx \ dy \ dt \\
&\leq \int_0^{T+\alpha} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |u(t,x) - u^0(x)|^2 \partial_t\varphi(t,x,y) \ dx \ dy \ dt \\
&+ \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |u^0(x) - u^0(y)|^2 \xi(x)\rho_\epsilon(x-y) \ dx \ dy \\
&- 2\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |u^0(x) - u^0(y)|\xi(x)\rho_\epsilon(x-y) \int_0^{T+\alpha} |u(t,x) - u^0(x)||\eta'_\alpha(t)| \ dt \ dx \ dy.
\end{aligned}$$

since $\|\eta'_\alpha\|_{L^1(0,T+\alpha)} = 1$. On the other hand, in inequality (6), take $\kappa = u^0(y)$ and the test function φ defined above and integrate it with respect to y; this gives

$$\begin{aligned}
&\int_0^{T+\alpha} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |u(t,x) - u^0(y)|^2 \partial_t\varphi(t,x,y) \ dx \ dy \ dt \\
&\qquad \geq -\int_0^{T+\alpha} \mathcal{R}_{\alpha,\epsilon}(t) \ dt - \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |u^0(x) - u^0(y)|^2 \xi(x)\rho_\epsilon(x-y) \ dx \ dy
\end{aligned}$$

where $\mathcal{R}_\epsilon \in L^1(0, T+\alpha)$ involves the remaining terms due to the space derivatives. The latter two inequalities can be combined, then letting α tend to 0 leads to

$$\begin{aligned}
&\int_{\mathbb{R}^d} |u(T,x) - u^0(x)|^2 \xi(x) \rho_\epsilon(x-y)\ dx \\
&\leq 2 \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |u^0(x) - u^0(y)|^2 \xi(x) \rho_\epsilon(x-y)\ dx\ dy + \int_0^{T+\alpha} \mathcal{R}_{0,\epsilon}(t)\ dt \\
&- 2 \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |u^0(x) - u^0(y)| \xi(x) \rho_\epsilon(x-y) |u(T,x) - u^0(x)|\ dx\ dy
\end{aligned}$$

Now, first take the limit $T \to 0^+$ in order to make the term with $\mathcal{R}_{0,\epsilon}$ vanish and, secondly, take the limit $\epsilon \to 0$ which, using the continuity of the translation operator, provides (26). □

Lemma 8. *Let u and $\widetilde{u}$ be two weak constrained solutions (in the sense of Definition 2) with initial data $u^0 \in L^2(\mathbb{R}^d)$ and $\widetilde{u}^0 \in L^2(\mathbb{R}^d)$. One has the Kato inequality: for all $\varphi \in \mathcal{D}_T^+$,*

$$\begin{aligned}
\int_{\mathbb{R}_+}\int_{\mathbb{R}^d} [|u-\widetilde{u}|^2 \partial_t \varphi + \sum_{i=1}^d \langle u - \widetilde{u}, A_i(u-\widetilde{u})\rangle\, \partial_{x_i}\varphi]\ dx\ dt \\
+ \int_{\mathbb{R}} |u^0 - \widetilde{u}^0|^2 \varphi(0,x)\ dx \geq 0. \quad (27)
\end{aligned}$$

Proof. The proof relies on the classical doubling of variables of Kruzhkov [14]. We consider (t,x) and (s,y) in $\mathbb{R}_+ \times \mathbb{R}^d$ and take a non-negative test function $\psi(t,x,s,y) \in \mathcal{C}_0^1((\mathbb{R}_+ \times \mathbb{R}^d)^2)$. We use inequality (6) for $u(t,x)$ (respectively $\tilde{u}(s,y)$), taking $\kappa = \tilde{u}(s,y)$ (resp. $\kappa = u(t,x)$) and $\varphi = \psi$ and integrate it with respect to $(s,y) \in \mathbb{R}_+ \times \mathbb{R}^d$ (resp. $(t,x) \in \mathbb{R}_+ \times \mathbb{R}^d$). After summation, it yields

$$\begin{aligned}
&\int_{\mathbb{R}_+}\int_{\mathbb{R}^d}\int_{\mathbb{R}_+}\int_{\mathbb{R}^d} \Big[|u(t,x) - \tilde{u}(s,y)|^2 (\partial_t + \partial_s)\psi(t,x,s,y) \\
&+ \sum_{i=1}^d \langle u(t,x) - \tilde{u}(s,y), A_i(u(t,x) - \tilde{u}(s,y))\rangle\, (\partial_{x_i} + \partial_{y_i})\psi(t,x,s,y) \Big]\ dx\, dt\, dy\, ds \\
&+ \int_{\mathbb{R}_+}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |(u(0,x) - \tilde{u}(s,y)|^2\ \psi(0,x,s,y)\ dx\ dy\ ds \\
&+ \int_{\mathbb{R}_+}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |(u(t,x) - \tilde{u}(0,y)|^2\ \psi(t,x,0,y)\ dy\ dx\ dt \geq 0. \quad (28)
\end{aligned}$$

Consider now $\Phi \in \mathcal{C}_0^1(\mathbb{R}_+ \times \mathbb{R}^d)$ and denote its partial derivatives $\partial_1 \Phi$ and $\partial_{z_j}\Phi$, $j \in \{1, ..., d\}$. For $i = \{1, ..., d\}$, let $\zeta_i \in \mathcal{C}^1(\mathbb{R}^i)$ be a function satisfying

$$\zeta_i(z) = \zeta_i(-z) \geq 0 \text{ for all } z \in \mathbb{R}^i,\ \text{supp}\zeta_i \subset [-1,1]^i,\ \text{and } \int_{\mathbb{R}^i} \zeta_i(z)\ dz = 1.$$

For $i = \{1, ..., d\}$ and $\epsilon > 0$, define the functions $\zeta_{i,\epsilon} \in \mathcal{C}^1(\mathbb{R}^i)$ by

$$\zeta_{i,\epsilon}(z) = \zeta_i(z/\epsilon)/\epsilon^i \text{ for all } z \in \mathbb{R}^i.$$

We set $\psi(t,x,s,y) = \Phi((t+s)/2, (x+y)/2)\zeta_{1,\epsilon}((t-s)/2)\zeta_{d,\epsilon}((x-y)/2)$, thus ψ satisfies

$$(\partial_t + \partial_s)\psi = \partial_1\Phi((t+s)/2, (x+y)/2)\ \zeta_{1,\epsilon}((t-s)/2)\ \zeta_{d,\epsilon}((x-y)/2),$$
$$(\partial_{x_i} + \partial_{y_i})\psi = \partial_{z_i}\Phi((t+s)/2, (x+y)/2)\ \zeta_{1,\epsilon}((t-s)/2)\ \zeta_{d,\epsilon}((x-y)/2).$$

Now, we use these identities in (28) and let ϵ go to 0, which leads to (27) due to the continuity at $t = 0$ provided by Proposition 7. □

We are now able to formulate the famous Kruzhkov's comparison result [14, 12].

Lemma 9. *Let $L = \max_i \rho(A_i)$ ($\rho(A)$ denotes the spectral radius of the matrix A). Let $u \in L^2([0,T] \times \mathbb{R}^d, K))$ and $\widetilde{u} \in L^2([0,T] \times \mathbb{R}^d, K))$ be two weak constrained solutions to (1) with initial data $u^0 \in L^2(\mathbb{R}^d, K)$ and $\widetilde{u}^0 \in L^2(\mathbb{R}^d, K)$. For all $T > 0$ and $r > 0$, one has the inequality*

$$\int_0^T \|u(T,\cdot) - \widetilde{u}(T,\cdot)\|^2_{L^2(B(0,r))}\, dt \le T \left\|u^0 - \widetilde{u}^0\right\|^2_{L^2(B(0,r+LT))}. \tag{29}$$

Continuous in time solutions actually satisfy a pointwise (in time) comparison principle: let $u \in \mathcal{C}([0,T], L^2(\mathbb{R}^d, K))$ and $\widetilde{u} \in \mathcal{C}([0,T], L^2(\mathbb{R}^d, K))$ be two weak constrained solutions to (1) with initial data $u^0 \in L^2(\mathbb{R}^d, K)$ and $\widetilde{u}^0 \in L^2(\mathbb{R}^d, K)$. For all $T > 0$ and $r > 0$, one has the inequality

$$\|u(T,\cdot) - \widetilde{u}(T,\cdot)\|_{L^2(B(0,r))} \le \left\|u^0 - \widetilde{u}^0\right\|_{L^2(B(0,r+LT))}. \tag{30}$$

Proof. Consider the function φ defined as

$$\varphi(t,x) = \begin{cases} \frac{T-t}{T} + \frac{r-|x|}{dLT} \text{ if } t \in [0,T] \text{ and } r < |x| < r + dL(T-t), \\ \frac{T-t}{T} \text{ if } t \in [0,T] \text{ and } x \in B(0,r), \\ 0 \text{ else} \end{cases}$$

as a test function in inequality (27) (φ is continuous and piecewise $\mathcal{C}^\infty$, so it can be taken as a test function). Computing $\partial_t\varphi$ and $\partial_{x_i}\varphi$, we see that

$$|u - \widetilde{u}|^2\partial_t\varphi + \sum_{i=1}^d \langle u - \widetilde{u}, A_i(u - \widetilde{u})\rangle\, \partial_{x_i}\varphi \le 0 \quad a.e.,$$

so that a consequence of inequality (27) is

$$-\int_0^T \int_{B(0,r)} [|u - \widetilde{u}|^2\partial_t\varphi + \sum_{i=1}^d \langle u - \widetilde{u}, A_i(u - \widetilde{u})\rangle\, \partial_{x_i}\varphi]\ dx\ dt \le \int_{\mathbb{R}} |u^0 - \widetilde{u}^0|^2\varphi(0,x)\ dx. \tag{31}$$

Now, remarking that in the domain of integration of the left-hand side above, $\partial_t \varphi = 1/T$ and $\partial_{x_i} \varphi = 0$ we obtain

$$-\frac{1}{T}\int_0^T \int_{B(0,r)} [|u - \widetilde{u}|^2 \leq \int_{\mathbb{R}} |u^0 - \widetilde{u}^0|^2 \varphi(0,x)\ dx$$

and finally, because $\varphi(0,\cdot) \leq 1$ and has its support in $B(0, r + dLT)$,

$$-\frac{1}{T}\int_0^T \int_{B(0,r)} [|u - \widetilde{u}|^2 \leq \int_{B(0,r+dLT)} |u^0 - \widetilde{u}^0|^2 \varphi(0,x)\ dx,$$

which is the first inequality announced.

The second one, concerning the continuous in time solutions, is let to the reader (it is classical, cf. [14], and can be proven with another well chosen test function).

□

Lemma 10. *Let $u \in \mathcal{C}([0,T], L^2(\mathbb{R}^d, K))$ be a continuous in time weak constrained solution with $u^0 \in H^1(\mathbb{R}^d)$. Then $u \in L^\infty([0,T], H^1(\mathbb{R}^d, K))$ with*

$$\|\nabla u(t)\|_{L^2(\mathbb{R}^d)} \leq \|\nabla u^0\|_{L^2(\mathbb{R}^d)}, \quad 0 < t \leq T. \tag{32}$$

Proof. Let $a \in \mathbb{R}^d$ be any vector. We set $u_a^0(x) = \widetilde{u}^0(x) = u^0(x-a)$. Since weak constrained solutions remains weak constrained solutions after translation then $\widetilde{u}(t,\cdot) = u(t,\cdot - a)$ is a weak constrained solution for this translated initial condition. Using inequality (30) for $r = \infty$ we obtain

$$\|u(t,\cdot) - u_a(t,\cdot)\|_{L^2(\mathbb{R}^d)} \leq \left\|u^0 - u_a^0\right\|_{L^2(\mathbb{R}^d)} \leq |a| \left\|\nabla u^0\right\|_{L^2(\mathbb{R}^d)}.$$

It is true for all vector $a \in \mathbb{R}^d$, therefore $u(t,\cdot) \in H^1(\mathbb{R}^d, K)$ satisfies (32). □

4 Existence

To prove the existence of a solution, we construct a simple numerical scheme and we prove the convergence of the numerical solutions to the weak constrained solution of (1) as the time and space steps tend to 0. For simplicity, the method is based on a splitting strategy between a PDE step and a projection step, and the mesh is Cartesian (for the analysis of the first step of the following algorithm on unstructured meshes, see [29] and [13]).

4.1 A numerical method

Let $\Delta x > 0$. Let the reference cell Ω_0 be d-dimensional hypercube

$$\Omega_0 = (0, \Delta x)^d.$$

For $i = \{1, ..., d\}$, let $\mathbf{T}_i$ be the translation operator in the ith direction,

$$\mathbf{T}_i u(\cdot) = u(\cdot - \Delta x e_i),$$

with e_i the ith vector of the canonical basis of $\mathbb{R}^d$. Then, for every $j = (j_i)_{i=1}^d \in \mathbb{Z}^d$, consider the translated cell

$$\Omega_j = \prod_{i=1}^{d} (j_i \Delta x, (j_i + 1)\Delta x) = \left(\prod_{i=1}^{d} \mathbf{T}_i^{j_i} \right) \Omega_0.$$

The set of cells $(\Omega_j)_{j \in \mathbb{Z}^d}$ is the mesh used for the algorithm and one has

$$\mathbb{R}^d = \cup_{j \in \mathbb{Z}^d} \overline{\Omega_j}, \quad \Omega_j \cap \Omega_k = \emptyset \text{ for } j \neq k.$$

The numerical approximate solution can be defined as constant in each cell at each time step (see the beginning of Section 4.3). In this case it is defined almost everywhere in the whole space $\mathbb{R}^d$ as

$$u_{\Delta x}^n(x) = u_j^n \text{ for } x \in \Omega_j \tag{33}$$

at time step n. Notice that by construction $\|u_{\Delta x}^n\|_{L^2(\mathbb{R}^d)}^2 = \Delta x^d \sum_{j \in \mathbb{Z}^d} \left|u_j^n\right|^2$, where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^n$.

We consider the following method where $\Delta t > 0$ is the time step.

Initialization: The numerical solution is classically defined at time step $n = 0$ by the mean value

$$u_j^0 = \frac{1}{|\Omega_j|} \int_{\Omega_j} u^0(x) dx \quad \forall j \in \mathbb{Z}^d. \tag{34}$$

So the whole task is to define U^{n+1} from U^n in the time loop.

Time iterate - Step 1 (Rusanov scheme): Advance in time with the discrete scheme

$$\frac{U^{n+1/2} - U^n}{\Delta t} + \sum_{i=1}^{d} A_i \frac{\mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x} U^n + c \sum_{i=1}^{d} \frac{2\mathbf{I} - \mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x} U^n = 0 \tag{35}$$

where the value of the coefficient c is chosen such that

$$c \geq \max_i \left(\rho(A_i) \right). \tag{36}$$

Notice that the first stage of time step can be rewritten in an explicit form $U^{n+1/2} = \mathbf{M} U^n$ with the iteration operator given by

$$\mathbf{M} = \left(1 - \frac{cd\Delta t}{\Delta x}\right) \mathbf{I} + \frac{\Delta t}{2\Delta x} \sum_{i=1}^{d} \left(c\mathbf{I} - A_i\right) \mathbf{T}_i^{-1} + \frac{\Delta t}{2\Delta x} \sum_j \left(c\mathbf{I} + A_j\right) \mathbf{T}_j.$$

Time iterate - Step 2: Apply the constraint exactly in each cell:

$$u_j^{n+1} = P_K \left(u_j^{n+1/2} \right) \quad \forall j.$$

Then we loop over the time step $\Delta t > 0$. As usual stability of the algorithm will be proven under a Courant-Friedrichs-Lewy (CFL) condition. We will show that in our case this CFL condition writes

$$cd\Delta t \leq \Delta x. \tag{37}$$

Therefore we assume that the parameters of the algorithm satisfy (37) and (36).

4.2 A priori estimates

Now we prove various a priori estimates. We begin with a standard inequality for the first step of the time iterate.

Let us define the local L^2 energy in the cell of the discrete solution at the beginning of the time iterate $e_j^n = \left|u_j^n\right|^2$, at the end of the first step of the time iterate $e_j^{n+1/2} = \left|u_j^{n+1/2}\right|^2$, and the local energy fluxes $(f_i^n)_j = \left\langle u_j^n, A_i u_j^n \right\rangle$ at the beginning of the time iterate.

Lemma 11. *Under the CFL condition (37), one has the inequality*

$$\frac{e^{n+1/2} - e^n}{\Delta t} + \sum_{i=1}^{d} \frac{\mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x} f_i^n + c \sum_{i=1}^{d} \frac{2\mathbf{I} - \mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x} e^n \leq 0 \tag{38}$$

where $X \leq Y$ means $X_j \leq Y_j$ for all $j \in \mathbb{Z}^d$.

Proof. By definition $U^{n+1/2} = \sum_{i=1}^{d} M_i U^n$ with

$$M_i = \left(\frac{1}{d} - \frac{c\Delta t}{\Delta x}\right)\mathbf{I} + \frac{\Delta t}{2\Delta x}\left(c\mathbf{I} - A_i\right)\mathbf{T}_i^{-1} + \frac{\Delta t}{2\Delta x}\left(c\mathbf{I} + A_i\right)\mathbf{T}_i.$$

• Let us prove the following upper bound (39) below. One has $(M_i U^n)_j = B_0 U_j^n + B_i V_j^n + C_i W_j^n$ with $V^n = \mathbf{T}_i^{-1} U^n$, $W^n = \mathbf{T}_i U^n$, $B_0 = \left(\frac{1}{d} - \frac{c\Delta t}{\Delta x}\right)\mathbf{I}$, $B_i = \frac{\Delta t}{2\Delta x}\left(c\mathbf{I} - A_i\right)$ and $C_i = \frac{\Delta t}{2\Delta x}\left(c\mathbf{I} + A_i\right)$. So

$$\left|(M_i U^n)_j\right|^2 = \left|B_0 u_j^n\right|^2 + \left|B_i v_j^n\right|^2 + \left|C_i w_j^n\right|^2$$

$$+2\left\langle B_0 u_j^n, B_i v_j^n \right\rangle + 2\left\langle B_0 u_j^n, C_i w_j^n \right\rangle + 2\left\langle B_i v_j^n, C_i w_j^n \right\rangle,$$

denoting $v_j^n = (V^n)_j$ and $w_j^n = (W^n)_j$. Due to the CFL condition (37) and to the inequality (36), all matrices B_0, B_i and C_i are symmetric and non-negative. On the other hand, they commute. So all products of these matrices are also non-negative and symmetric matrices. One can use the Cauchy-Schwarz inequality to show for example that

$$2\left\langle B_i v_j^n, C_i w_j^n \right\rangle \leq \left\langle C_i B_i v_j^n, v_j^n \right\rangle + \left\langle B_i C_i w_j^n, w_j^n \right\rangle$$

(and similar inequalities for similar expressions). Therefore

$$\begin{aligned}\left|(M_iU^n)_j\right|^2 &\leq \left|B_0u_j^n\right|^2+\left|B_iv_j^n\right|^2+\left|C_iw_j^n\right|^2\\ &\quad+\left\langle B_iB_0u_j^n,u_j^n\right\rangle+\left\langle B_0B_iv_j^n,v_j^n\right\rangle\\ &\quad+\left\langle C_iB_0u_j^n,u_j^n\right\rangle+\left\langle B_0C_iw_j^n,w_j^n\right\rangle\\ &\quad+\left\langle C_iB_iv_j^n,v_j^n\right\rangle+\left\langle B_iC_iw_j^n,w_j^n\right\rangle\\ &\leq \left\langle D_iB_0u_j^n,u_j^n\right\rangle+\left\langle D_iB_iv_j^n,v_j^n\right\rangle+\left\langle D_iC_iw_j^n,w_j^n\right\rangle\end{aligned}$$

with $D_i=B_0+B_i+C_i$. We notice that $D_i=\frac{1}{d}\mathbf{I}$ for all i. It means that

$$\left|(M_iU^n)_j\right|^2\leq\frac{1}{d}\left\langle B_0u_j^n,u_j^n\right\rangle+\frac{1}{d}\left\langle B_iv_j^n,v_j^n\right\rangle+\frac{1}{d}\left\langle C_iw_j^n,w_j^n\right\rangle. \tag{39}$$

• Therefore

$$\begin{aligned}e_j^{n+1/2}=\left|u_j^{n+1/2}\right|^2&\leq d\sum_{i=1}^d\left|(M_iU^n)_j\right|^2\\ &\leq\sum_{i=1}^d\left[\left\langle B_0u_j^n,u_j^n\right\rangle+\left\langle B_iv_j^n,v_j^n\right\rangle+\left\langle C_iw_j^n,w_j^n\right\rangle\right]\\ &\leq\left(1-\frac{cd\Delta t}{\Delta x}\right)\left|u_j^n\right|^2+\sum_{i=1}^d\left[\left\langle B_iv_j^n,v_j^n\right\rangle+\left\langle C_iw_j^n,w_j^n\right\rangle\right].\end{aligned}$$

By definition, one has the identities $\left\langle B_iv_j^n,v_j^n\right\rangle=\frac{c\Delta t}{2\Delta x}\left(\mathbf{T}_i^{-1}e^n\right)_j-\frac{\Delta t}{2\Delta x}\left(\mathbf{T}_i^{-1}f_i^n\right)_j$ and $\left\langle C_iw_j^n,w_j^n\right\rangle=\frac{c\Delta t}{2\Delta x}\left(\mathbf{T}_ie^n\right)_j+\frac{\Delta t}{2\Delta x}\left(\mathbf{T}_if_i^n\right)_j$. Finally we obtain

$$\begin{aligned}e_j^{n+1/2}&\leq\left(1-\frac{cd\Delta t}{\Delta x}\right)e_j^n\\ &+\sum_{i=1}^d\left[\frac{c\Delta t}{2\Delta x}\left(\mathbf{T}_i^{-1}e^n\right)_j-\frac{\Delta t}{2\Delta x}\left(\mathbf{T}_i^{-1}f_i^n\right)_j+\frac{c\Delta t}{2\Delta x}\left(\mathbf{T}_ie^n\right)_j+\frac{\Delta t}{2\Delta x}\left(\mathbf{T}_if_i^n\right)_j\right],\end{aligned}$$

which is nothing but the claim. □

Next we consider a test vector $\kappa\in K$ and we define

$$(e_\kappa^n)_j=\left|u_j^n-\kappa\right|^2 \tag{40}$$

and the local energy fluxes

$$(f_{i,\kappa}^n)_j=\left\langle u_j^n-\kappa,A_i\left(u_j^n-\kappa\right)\right\rangle.$$

These quantities correspond to the discrete analogues of what appears in the weak inequalities (6).

Corollary 12. *Under the CFL condition (37), one has the inequality for all* $\kappa \in K$

$$\frac{e_\kappa^{n+1} - e_\kappa^n}{\Delta t} + \sum_{i=1}^{d} \frac{\mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x} f_{i,\kappa}^n + c \sum_{i=1}^{d} \frac{2\mathbf{I} - \mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x} e_\kappa^n \leq 0. \tag{41}$$

Proof. Since the first step of the time iterate is invariant with respect to translations in the state space, one can generalize (38) and get for all $\kappa \in K$

$$\frac{e_\kappa^{n+1/2} - e_\kappa^n}{\Delta t} + \sum_{i=1}^{d} \frac{\mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x} f_{i,\kappa}^n + c \sum_{i=1}^{d} \frac{2\mathbf{I} - \mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x} e_\kappa^n \leq 0.$$

Here $\left(e_\kappa^{n+1/2}\right)_j = \left|u_j^{n+1/2} - \kappa\right|^2$. The second step of the algorithm is an L^2 contraction in $\mathbb{R}^d$, since it is a projection; so $e_\kappa^{n+1} \leq e_\kappa^{n+1/2}$. With the previous inequality it proves the claim. □

Remark 13. *Inequality (41) is connected to the classical* L^2 *stability of the scheme. Indeed one has*

$$\|u_{\Delta x}^n\|_{L^2(\mathbb{R}^d)} \leq \left\|u_{\Delta x}^0\right\|_{L^2(\mathbb{R}^d)}. \tag{42}$$

To prove this, we use inequality (41) with $\kappa = 0$*. Then we sum the components of* e_0^{n+1} *and* e_0^n *over all cells. Since the fluxes disappear then one gets* $\sum_j \left(e_0^{n+1}\right)_j \leq \sum_j \left(e_0^n\right)_j$ *which implies (42). Inequality (37) is the CFL condition for the classical* L^2 *stability.*

Lemma 14. *Consider two numerical solutions* U^n *and* V^n*. Then, under the CFL condition (37), one has the inequality*

$$\|U^n - V^n\|_{L^2(\mathbb{R}^d)} \leq \left\|U^0 - V^0\right\|_{L^2(\mathbb{R}^d)}. \tag{43}$$

Proof. Set $W^n = U^n - V^n$ and $W^{n+1/2} = U^{n+1/2} - V^{n+1/2}$. Since the first step of the time loop is a linear scheme then $W^{n+1/2} = \mathbf{M}W^n$. Since this first step is L^2 stable (38) then $\left\|W^{n+1/2}\right\|_{L^2(\mathbb{R}^d)} \leq |W^n\|_{L^2(\mathbb{R}^d)}$. With classical notations one has

$$\begin{aligned} W^{n+1} =& P_K U^{n+1/2} - P_K V^{n+1/2} \\ &\Rightarrow \|W^{n+1}\|_{L^2(\mathbb{R}^d)} \leq \|U^{n+1/2} - V^{n+1/2}\|_{L^2(\mathbb{R}^d)} = \|W^{n+1/2}\|_{L^2(\mathbb{R}^d)} \end{aligned}$$

since P_K is, in each cell, a projection onto the convex set K. So $\left\|W^{n+1}\right\|_{L^2(\mathbb{R}^d)} \leq \|W^n\|_{L^2(\mathbb{R}^d)}$. □

Lemma 15. *Assume* $u^0 \in H^1(\mathbb{R}^d, K)$*. Assume the CFL condition (37) holds. One has the space inequalities*

$$\|U^n - \mathbf{T}_i U^n\|_{L^2(\mathbb{R}^d)} \leq \Delta x \left\|u^0\right\|_{H^1(\mathbb{R}^d)} \quad \forall i \in \{1, ..., d\}. \tag{44}$$

The time inequality

$$\left\|U^{n+1}-U^n\right\|_{L^2(\mathbb{R}^d)} \leq C\Delta t \left\|u^0\right\|_{H^1(\mathbb{R}^d)} \tag{45}$$

also holds, for some constant $C>0$ only depending on the matrices A_i.

Proof. To prove the first inequality (44) we use (43) with $V^n = \mathbf{T}_i U^n$ and the well-known bound $\left\|U^0 - \mathbf{T}_i U^0\right\|_{L^2(\mathbb{R}^d)} \leq \Delta x \left\|u^0\right\|_{H^1(\mathbb{R}^d)}$.

To prove the second inequality, (45), we rely on the triangular inequality

$$\left\|U^{n+1}-U^n\right\|_{L^2(\mathbb{R}^d)} \leq \left\|U^{n+1}-U^{n+1/2}\right\|_{L^2(\mathbb{R}^d)} + \left\|U^{n+1/2}-U^n\right\|_{L^2(\mathbb{R}^d)}.$$

Considering (35), the increment $U^{n+1/2}-U^n$ is easily expressed as $\frac{\Delta t}{\Delta x}$ times a linear combination of terms that can all be bounded using (44). Therefore the following inequality holds for a convenient constant $C_1>0$

$$\left\|U^{n+1/2}-U^n\right\|_{L^2(\mathbb{R}^d)} \leq C_1\Delta t \left\|u^0\right\|_{H^1(\mathbb{R}^d)}.$$

By construction, $u_j^n \in K$ for all $j \in \mathbb{Z}^d$. Therefore

$$\left|u_j^{n+1/2}-u_j^{n+1}\right| = \left|u_j^{n+1/2} - P_K\left(u_j^{n+1/2}\right)\right| \leq \left|u_j^{n+1/2}-u_j^n\right|.$$

So

$$\left\|U^{n+1}-U^{n+1/2}\right\|_{L^2(\mathbb{R}^d)} \leq C_1\Delta t \left\|u^0\right\|_{H^1(\mathbb{R}^d)}.$$

Therefore (45) holds with $C=2C_1$. It ends the proof. □

4.3 Convergence to the weak solution

In this section, the existence result stated in Theorem 3 is obtained by proving that the sequence of numerical solutions converges (up to a subsequence) toward a solution of the constrained problem in the sense of Definition 2 as the space step Δx tends to 0. Here the ratio $\Delta t/\Delta x$ is fixed:

$$\Delta t = C_{CFL}\Delta x$$

(with $C_{CFL} \leq 1/(cd)$), so that the unique parameter of the numerical solution is Δx.

Before stating the main results of this part, we introduce the barycentric functions of the unit reference hypercube $S=[0,1]^d$. These barycentric functions enable us to construct representation functions of the numerical solution and to apply standard compactness results.

By definition, the d-dimensional hypercube has 2^d corners, each of which is denoted

$$\alpha = (\alpha_1, \cdots, \alpha_d) \quad \alpha_i = 0 \text{ or } 1, \quad 1 \leq i \leq d.$$

The barycentric functions are

$$\lambda_\alpha(x) = \Pi_{i=1}^d \left(1 - \alpha_i + (2\alpha_i - 1)x_i\right) \mathbf{1}_S(x), \quad x = (x_i)_{1\le i\le d} \in S.$$

By construction, $\lambda_\alpha(\beta) = \delta_{\alpha,\beta}$ for $\alpha, \beta \in \{0,1\}^d$. Starting from the numerical solution (33) we define a first function almost everywhere

$$u_{\Delta x}(t,x) = u_j^n \text{ for } x \in \Omega_j \text{ and } t_n < t < t_{n+1} \tag{46}$$

where $t_n = n\Delta t$. We define a second function almost everywhere

$$U_{\Delta x}(t,x) = \sum_{\alpha\in\{0,1\}^d} \lambda_\alpha\left(\frac{x - j\Delta x}{\Delta x}\right) u_{j+\alpha}^n, \quad x \in \Omega_j,\ t_n < t < t_{n+1}. \tag{47}$$

In the language of numerical analysis, $U_{\Delta x}$ is a Q^1 reconstruction of the numerical solution. Notice that $U_{\Delta x}$ is by construction continuous and affine in all canonical directions inside every cell.

Lemma 16. *There exists three positive constants c_1, c_2 and c_3 such that for a.e. $t > 0$*

$$\|U_{\Delta x}(t)\|_{L^2(\mathbb{R}^d)} \le c_1 \|u^0\|_{L^2(\mathbb{R}^d)}, \tag{48}$$

$$\|\nabla U_{\Delta x}(t)\|_{L^2(\mathbb{R}^d)} \le c_2 \|\nabla u^0\|_{L^2(\mathbb{R}^d)} \text{ if } u^0 \in H^1(\mathbb{R}^d), \tag{49}$$

and

$$\|U_{\Delta x}(t) - u_{\Delta x}(t)\|_{L^2(\mathbb{R}^d)} \le \left(c_3 \|\nabla u^0\|_{L^2(\mathbb{R}^d)}\right) \Delta x \text{ if } u^0 \in H^1(\mathbb{R}^d). \tag{50}$$

Proof. Let us consider the cell

$$j\Delta x + S\Delta x = \left\{x = (x_i),\ j_i\Delta x < x_i < (j_i+1)\Delta x\right\}, \quad j = (j_i)_{1\le i\le d}.$$

By construction, the barycentric functions are such that $0 \le \lambda_\alpha \le 1$, so one deduces

$$\int_{j\Delta x + S\Delta x} |U_{\Delta x}(t,x)|^2 \, dx \le \frac{c_1}{2^d} \sum_{\alpha\in\{0,1\}^d} \left|u_{j+\alpha}^n\right|^2, \quad t_n < t < t_{n+1}.$$

After summation over all j and the use of (42), the inequality (48) is obtained.

From (47) one gets for $x \in \Omega_j$ and $t_n < t < t_{n+1}$

$$\nabla U_{\Delta x}(t,x) = \frac{1}{\Delta x} \sum_{\alpha\in\{0,1\}^d} \nabla\lambda_\alpha\left(\frac{x - j\Delta x}{\Delta x}\right) u_{j+\alpha}^n.$$

Since $\sum_\alpha \lambda_\alpha = 1$ by definition, then $\sum_\alpha \nabla\lambda_\alpha = 0$. So one has also

$$\nabla U_{\Delta x}(t,x) = \frac{1}{\Delta x} \sum_{\alpha\in\{0,1\}^d} \nabla\lambda_\alpha\left(\frac{x - j\Delta x}{\Delta x}\right) \left(u_{j+\alpha}^n - u_j^n\right)$$

for $x \in \Omega_j$ and $t_n < t < t_{n+1}$. One obtains the bound

$$\int_{j\Delta x + S\Delta x} |\nabla U_{\Delta x}(t,x)|^2 \, dx \leq \frac{\tilde{c}_2}{2^d} \sum_{\alpha \in \{0,1\}^d} \left| u^n_{j+\alpha} - u^n_j \right|^2, \; t_n < t < t_{n+1}.$$

Using now the inequality (44) we obtain the inequality (49) with eventually a greater constant $c_2 \geq \tilde{c}_2$.

Finally we notice that

$$U_{\Delta x}(t,x) - u_{\Delta x}(t,x) = \sum_{\alpha \in \{0,1\}^d} \lambda_\alpha \left(\frac{x - j\Delta x}{\Delta x} \right) \left(u^n_{j+\alpha} - u^n_j \right)$$

for $x \in \Omega_j$ and $t_n < t < t_{n+1}$, from which we deduce (50). □

Lemma 17. *Assume $u^0 \in H^1(\mathbb{R}^d, K)$ has a compact support. Let $T > 0$. Then there exists a weak solution $u \in \mathcal{C}([0,T], L^2(\mathbb{R}^d, K)) \cap L^\infty([0,T], H^1(\mathbb{R}^d, K))$ to (1) in the sense of Definition 2. Moreover $u(t,\cdot)$ has a compact support in space for every $t \in [0,T]$.*

Proof. The proof consists in two stages: first we let Δx go to zero and we extract a converging subsequence; second we show that the limit is indeed a weak solution.

Let us denote by C the support of $u^0 \in H^1(\mathbb{R}^d, K)$: C is bounded by hypothesis. For any $\Delta x > 0$, let $U_{\Delta x}$ be the piecewise affine by direction function (47) defined by the scheme with this initial condition u^0. Let $F = \{U_{\Delta x}, \; 0 < \Delta x < 1\}$. For all $t \in [0,T]$ and all $\Delta x \in (0,1]$, the support of $U_{\Delta x}(t)$ is included in $C_T = C + [-cT - 1, cT + 1]^d$ where c satisfies (36), thanks to the (crucial) finite propagation velocity property of the numerical scheme.

From inequality (48),

$$F \subset L^\infty([0,T], L^2(C_T)).$$

From inequality (49),

$$F \subset L^\infty([0,T], H^1(C_T)),$$

thus $F \subset L^1_{loc}([0,T], H^1(C_T))$. From inequality (45), for every $h > 0$, denoting by τ_h the translation operator with length h in the t direction,

$$||\tau_h f - f||_{L^\infty([0,T], L^2(C_T))} \leq Ch ||u^0||_{H^1(\mathbb{R}^d)}$$

for every $f \in F$, thus $||\tau_h f - f||_{L^\infty([0,T], L^2(C_T))}$ converges to 0 uniformly for $f \in F$. Finally, since C_T is bounded, the imbedding $H^1(C_T) \subset L^2(C_T)$ is compact. Therefore from a classical compactness result (Theorem 3 of [26] for instance), there exists a sequence $(\Delta x_n)_{n \in \mathbb{N}}$ converging to 0 such that $U_{\Delta x_n}$ converges toward $u \in \mathcal{C}([0,T], L^2(C_T))$. One also has by construction that $u \in L^\infty([0,T], H^1(\mathbb{R}^d, K))$ and $\mathrm{supp}(u) \subset C_T$.

Due to (50), $u_{\Delta x}$ also converges to $u \in \mathcal{C}([0,T], L^2(C_T))$. We will use this property to prove that u is a solution of the problem. The method is in the

spirit of the Lax-Wendroff theorem. Let us consider a test function $\varphi \in \mathcal{D}_T^+$ and set $\varphi_j^n = \varphi(n\Delta t, j\Delta x)$. For a test vector $\kappa \in K$ we define e_κ as in 40, multiply inequality (41) by $\Delta x^d \Delta t \varphi_j^n$, and take the sum

$$\Delta x^d \Delta t \sum_{j\in\mathbb{Z}^d, n\in\mathbb{N}} \left(\frac{e_\kappa^{n+1} - e_\kappa^n}{\Delta t} + \sum_{i=1}^d \left(\frac{\mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x} f_{i,\kappa}^n + c\frac{2\mathbf{I} - \mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x} e_\kappa^n \right) \right)_j \varphi_j^n \leq 0. \quad (51)$$

Successive uses of the Abel rule lead to

$$\begin{aligned} &- \Delta x^d \Delta t \sum_{j\in\mathbb{Z}^d, n\in\mathbb{N}^*} e_\kappa^n \frac{\varphi_j^n - \varphi_j^{n-1}}{\Delta t} - \Delta x^d \sum_{j\in\mathbb{Z}^d} e_\kappa^0 \varphi_j^0 \\ &- \Delta x^d \Delta t \sum_{j\in\mathbb{Z}^d, n\in\mathbb{N}} \sum_{i=1}^d \left(f_{i,\kappa}^n\right)_j \left(\frac{\mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x} \varphi^n \right)_j \\ &- \left(\Delta x^d \Delta t \sum_{j\in\mathbb{Z}^d, n\in\mathbb{N}} \sum_{i=1}^d (e_\kappa^n)_j \left(\frac{2\mathbf{I} - \mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x^2} \varphi^n \right)_j \right) c\Delta x \leq 0. \end{aligned} \quad (52)$$

Now we consider the limit of this inequality in the regime $\Delta x \to 0$ (recall that the ratio $\Delta t/\Delta x$ is fixed). Since $e_\kappa \to |u-\kappa|^2$ and $e_\kappa^0 \to |u^0 - \kappa|^2$ in L^1,

$$\Delta x^d \Delta t \sum_{j\in\mathbb{Z}^d, n\in\mathbb{N}^*} e_\kappa^n \frac{\varphi_j^n - \varphi_j^{n-1}}{\Delta t} \longrightarrow \int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \frac{|u-\kappa|^2}{2} \partial_t \varphi$$

and

$$\Delta x^d \sum_{j\in\mathbb{Z}^d} e_\kappa^0 \varphi_j^0 \longrightarrow \int_{\mathbb{R}^d} \frac{|u^0 - \kappa|^2}{2} \varphi^0.$$

Similarly,

$$\begin{aligned} &\Delta x^d \Delta t \sum_{j\in\mathbb{Z}^d, n\in\mathbb{N}} \left(f_{i,\kappa}^n\right)_j \left(\frac{\mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x} \varphi^n \right)_j \\ &\longrightarrow \int_{\mathbb{R}^d} \int_{\mathbb{R}_+} \langle u - \kappa, A_i(u-\kappa) \rangle \, \partial_{x_i} \varphi dt dx. \end{aligned}$$

Finally

$$\left(\Delta x^d \Delta t \sum_{j\in\mathbb{Z}^d, n\in\mathbb{N}} (e_\kappa^n)_j \left(\frac{2\mathbf{I} - \mathbf{T}_i^{-1} - \mathbf{T}_i}{2\Delta x^2} \varphi^n \right)_j \right) c\Delta x \to 0.$$

Therefore passing to the limit in (51) we get the existence of a weak solution. □

Proof of Theorem 3. Let $\left(u_k^0\right)_{k\in\mathbb{N}}$ be a sequence of functions in $H^1(\mathbb{R}^d, K)$ with compact support that converges toward u^0 in $L^2(\mathbb{R}^d, K)$. Thanks to Lemma 17, for every $k \in \mathbb{N}$, there exists an associated weak constrained solution $u_k \in \mathcal{C}([0,T], L^2(\mathbb{R}^d, K))$. For every $k, m \in \mathbb{N}$, one has, from estimate (43) (that remains true at the limit for the exact solution),

$$\|u_k(t,\cdot) - u_m(t,\cdot)\|_{L^2(\mathbb{R}^d)} \leq \|u_k^0 - u_m^0\|_{L^2(\mathbb{R}^d)}.$$

Thus, for every $t \in [0,T]$, $\left(u_k(t,\cdot)\right)_{k\in\mathbb{N}}$ converges in L^2 to a function $u(t,\cdot)$. Furthermore, letting m go to infinity leads to

$$\|u_k(t,\cdot) - u(t,\cdot)\|_{L^2(\mathbb{R}^d)} \leq \|u_k^0 - u^0\|_{L^2(\mathbb{R}^d)}.$$

It can be shown, in the same manner as in the proof of Lemma 17, that u is a solution to the problem with initial condition u^0. Let us show that $u \in \mathcal{C}([0,T], L^2(\mathbb{R}^d, K))$, that is to say that u is continuous with respect to time. For every $t, s \in [0,T]$ one has

$$\|u(t,\cdot) - u(s,\cdot)\|_{L^2(\mathbb{R}^d)}$$

$$\leq \|u(t,\cdot) - u_k(t,\cdot)\|_{L^2(\mathbb{R}^d)} + \|u_k(t,\cdot) - u_k(s,\cdot)\|_{L^2(\mathbb{R}^d)} + \|u_k(s,\cdot) - u(s,\cdot)\|_{L^2(\mathbb{R}^d)}$$

$$\leq 2\|u_k^0 - u^0\|_{L^2(\mathbb{R}^d)} + \|u_k(t,\cdot) - u_k(s,\cdot)\|_{L^2(\mathbb{R}^d)}.$$

As u_k^0 converges to u^0 and u^k is continuous in time, this inequality shows that for every $\varepsilon > 0$ there exists η such that if $|t-s| \leq \eta$, $\|u(t,\cdot) - u(s,\cdot)\|_{L^2(\mathbb{R}^d)} \leq \varepsilon$. Thus we have a solution to the problem in $\mathcal{C}([0,T], L^2(\mathbb{R}^d, K))$. Lemma 9 shows that problem (1) has at most one solution in the sense of Definition 2. The theorem is proven.

If the initial condition lies in H^1, the solution lies in $L^\infty([0,T], H^1(\mathbb{R}^d, K))$ as a consequence of inequality (32). In particular, this shows that the presence of a constraint in the problem does not produce discontinuities in the solution. □

5 Discontinuous solutions in dimension 1

In this section we consider discontinuous weak solutions, in dimension $d = 1$ for simplicity.

Let us consider a weak solution u defined by

$$u(t,x) = \begin{cases} u_L \in K \text{ for } x < \sigma t, \\ u_R \in K \text{ for } x > \sigma t \end{cases} \tag{53}$$

for $t \geq 0$ and $\sigma \in \mathbb{R}$. The weak inequality writes

$$\int_0^T \int_{\mathbb{R}} \left[\frac{|u-\kappa|^2}{2} \partial_t + \frac{\langle u-\kappa, A(u-\kappa)\rangle}{2} \partial_x \right] \varphi \, dx \, dt + \int_{\mathbb{R}} \frac{|u^0-\kappa|^2}{2} \varphi(0,x) \, dx \geq 0 \tag{54}$$

for all $\kappa \in K$ and for all $\varphi \in \mathcal{D}_T^+$.

The main question is to characterize all admissible triplets (σ, u_L, u_R) and to determine if these triplets are solutions, or not, of the unconstrained Rankine-Hugoniot relations: $\sigma\,(u_R - u_L) = A\,(u_R - u_L)$.

Lemma 18. *A discontinuous solution of type (53) is a weak solution if and only if*

$$\begin{aligned} -\,\sigma\left[|u_R-\kappa|^2-|u_L-\kappa|^2\right] \\ +\left[\langle u_R-\kappa, A(u_R-\kappa)\rangle - \langle u_L-\kappa, A(u_L-\kappa)\rangle\right] \le 0 \quad \forall \kappa \in K. \end{aligned} \tag{55}$$

These relations are called weak Rankine-Hugoniot relations in [8].

Definition 19. *The closed convex set K is strictly convex if and only if*

$$\textit{for all } \kappa, \kappa' \in K, \ \textit{for all } \alpha \in (0,1), \quad \alpha\kappa + (1-\alpha)\kappa' \in \textit{Int}(K).$$

Theorem 20. *Assume that K is strictly convex. Then a solution to the weak Rankine-Hugoniot relations (55) is a solution of the classical Rankine-Hugoniot relations*

$$-\sigma(u_R - u_L) + A(u_r - u_L) = 0.$$

Proof. After some algebra, (55) rewrites

$$\left\langle -\sigma(u_R-u_L) + A(u_R-u_L), \kappa - \frac{1}{2}(u_R+u_L) \right\rangle \le 0 \quad \forall \kappa \in K.$$

We plug $\kappa = \frac{1}{2}(u_R+u_L) + w$ in (55).

The strict convexity of K implies that there exists $\varepsilon > 0$ such that for all $w \in B(0,\varepsilon)$, $\kappa = \frac{1}{2}(u_R+u_L) + w \in K$. Plugging this κ in (55) provides

$$\langle -\sigma(u_R-u_L) + A(u_R-u_L), w\rangle \le 0 \quad \forall w \in B(0,\varepsilon).$$

Therefore $-\sigma(u_R-u_L) + A(u_R-u_L) = 0$ and the claim is proven. □

6 Numerical examples

In order to illustrate the previous theory, we compute the numerical solution for the wave system (7) in dimension one with the convex $K = [0,1]^2$ (subsection 6.1), and for the isotropic strain hardening system

$$\begin{cases} \partial_t u - \partial_x \sigma = 0, \\ \partial_t \sigma - \partial_x u = 0, \\ \partial_t \gamma = 0, \\ |\sigma| + \gamma \le 1, \end{cases}$$

(subsection 6.2). To simplify the interpretation of the numerical solutions, the Courant number is exactly 1: $\Delta t = \Delta x$. In this case the dissipation of the scheme vanishes, so the profiles remain sharp.

6.1 Wave system

The domain is $\Omega = [0, 1]$ with periodic boundary conditions. The initial condition is the discontinuous profile pictured in figure 1. The solution at $t = 0.2$ and $t = 0.4$ is displayed in figures 2 and 3, with a comparison with the unconstrained solution. The phase diagram of the numerical at times $t = 0.2$ and $t = 0.4$ is plotted in figure 4. On this figure we also plot the domain K which is a square, together with the phase diagram of the unconstrained solution. At time $t = 0.2$ some parts of the unconstrained solution are clearly outside of K. We observe that the numerical constrained solution satisfies the constraint.

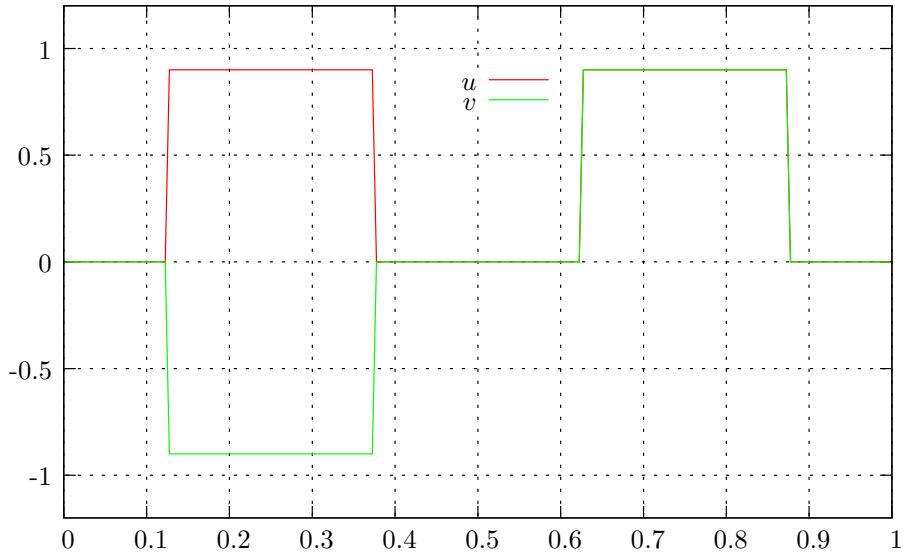

Figure 1: Initial condition

6.2 Strain hardening

We solve numerically the problem (22) with the constraint (24), $g(\gamma) = \gamma$ and $k = 1$). The spatial domain is $[0, 1]$. We take as initial condition $u(0, x) = \sigma(0, x) = \gamma(0, x) = 0$ for all $x \in [0, 1]$. At the left boundary, we impose a given pulsation:

$$(u - \sigma)(t, 0) = 3 \sin(12\pi t), \quad t \in \mathbb{R}_+.$$

The right boundary is endowed with the "transparent" condition

$$\partial_x(u + \sigma)(t, 1) = 0.$$

We observe on figure 4 that the numerical solution is such that γ decreases as the time advances. It means that the size of the static domain $|\sigma|_{\max} = k - g(\gamma)$

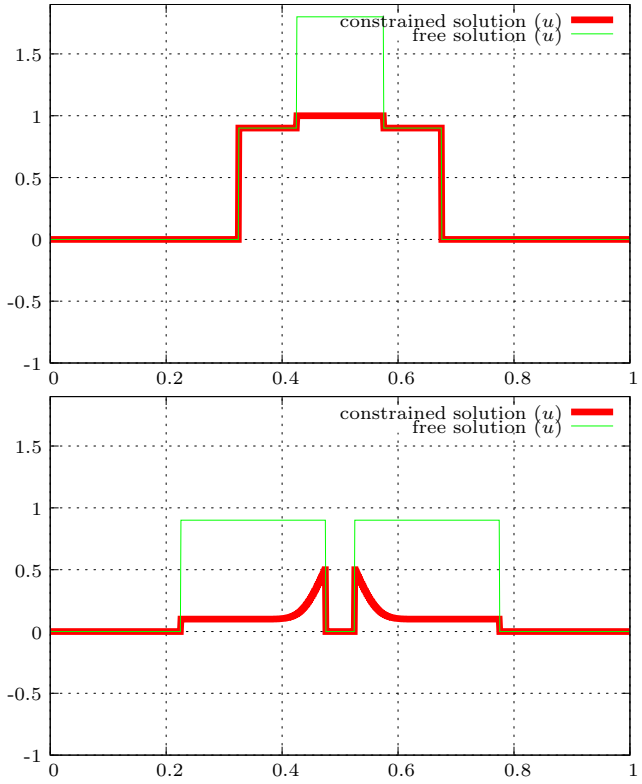

Figure 2: Numerical solution u at times $t = 0.2$ (top) and $t = 0.4$ (bottom). On both pictures we compare the constrained solution (bold line in red) and the unconstrained solution (thin line in green).

increases if the strain hardening parameter γ decreases. That is to say, that the constraint is less and less severe for the stress variable σ. This is all the more visible on figure 5 where we plot the function

$$t \mapsto (\gamma(t, 1/2), \sigma(t, 1/2)) .$$

A Convergence of the relaxation model

In this section, we give the main estimate that can be used to prove the convergence of the classical solutions of the relaxation model (2) towards the weak constrained solution of (1) when $\varepsilon \to 0$.

Lemma 21. *Assume $u^0 \in H^1(\mathbb{R}^d, K)$ and consider $u_\varepsilon \in L^2([0,T] \times \mathbb{R}^d)$ the weak solution of the relaxation problem (2). Therefore, for all $T > 0$ and all*

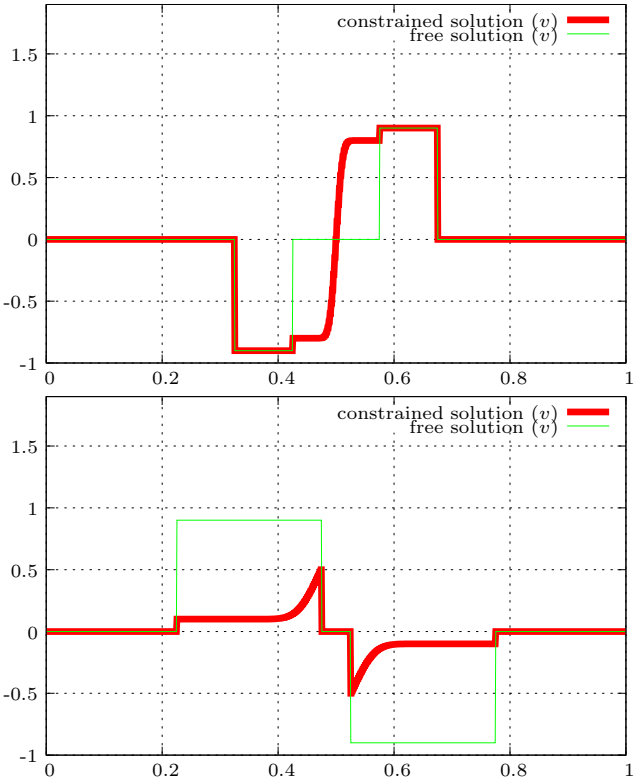

Figure 3: Numerical solution v at times $t = 0.2$ (top) and $t = 0.4$ (bottom). On both pictures we compare the constrained solution (bold line in red) and the unconstrained solution (thin line in green).

$r > 0$, there exists a positive constant C which only depends on T and r, such that

$$\|P_K u_\varepsilon - u_\varepsilon\|^2_{L^2((0,T)\times B(0,r))} \leq C\frac{\varepsilon}{2}|u^0|^2_{H^1(B(0,r+LT))} \tag{56}$$

where $|\cdot|$ denotes a semi-norm.

Proof. The weak solution u_ε satisfies for all constant vector $\kappa \in K$ and non-negative test function φ the equation

$$\int_{\mathbb{R}}\int_{\mathbb{R}^d}\left[|u_\varepsilon - \kappa|^2\partial_t + \sum_{i=1}^{d}\langle u_\varepsilon - \kappa, A_i(u_\varepsilon - \kappa)\rangle\,\partial_{x_i}\right]\varphi\,dx\,dt$$
$$+\int_{\mathbb{R}^d}|u^0_\varepsilon - \kappa|^2\varphi(0,x)\,dx + \frac{2}{\varepsilon}\int_{\mathbb{R}}\int_{\mathbb{R}^d} < u_\varepsilon - \kappa, P_K u_\varepsilon - u_\varepsilon > \varphi\,dx\,dt \geq 0. \tag{57}$$

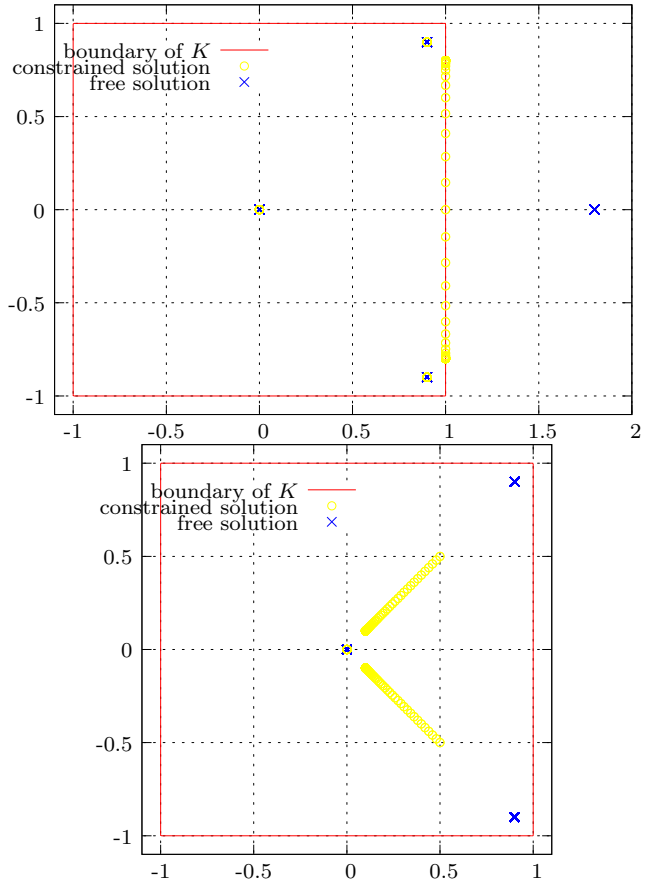

Figure 4: Phase diagram u, v at times $t = 0.2$ (top) and $t = 0.4$ (bottom). The phase diagram of the unconstrained solution is composed of 3 points (the solution is piecewise constant). The phase diagram of the constrained solution is inside the domain K which is a square.

By definition of P_K, we have $(u_\varepsilon - \kappa, P_K u_\varepsilon - u_\varepsilon) \leq -|u_\varepsilon - P_K u_\varepsilon|^2$, which yields

$$\int_{\mathbb{R}} \int_{\mathbb{R}^d} \left[|u_\varepsilon - \kappa|^2 \partial_t + \sum_{i=1}^{d} \langle u_\varepsilon - \kappa, A_i(u_\varepsilon - \kappa) \rangle \, \partial_{x_i} \right] \varphi \, dx \, dt$$
$$+ \int_{\mathbb{R}^d} |u_\varepsilon^0 - \kappa|^2 \varphi(0, x) \, dx - \frac{2}{\varepsilon} \int_{\mathbb{R}} \int_{\mathbb{R}^d} |P_K u_\varepsilon - u_\varepsilon|^2 \, \varphi \, dx \, dt \geq 0. \quad (58)$$

As in the Kruzhkov's uniqueness proof, take $\varphi(t, x) = \chi_\eta(t) \omega_\eta(t, x)$, $\eta > 0$,

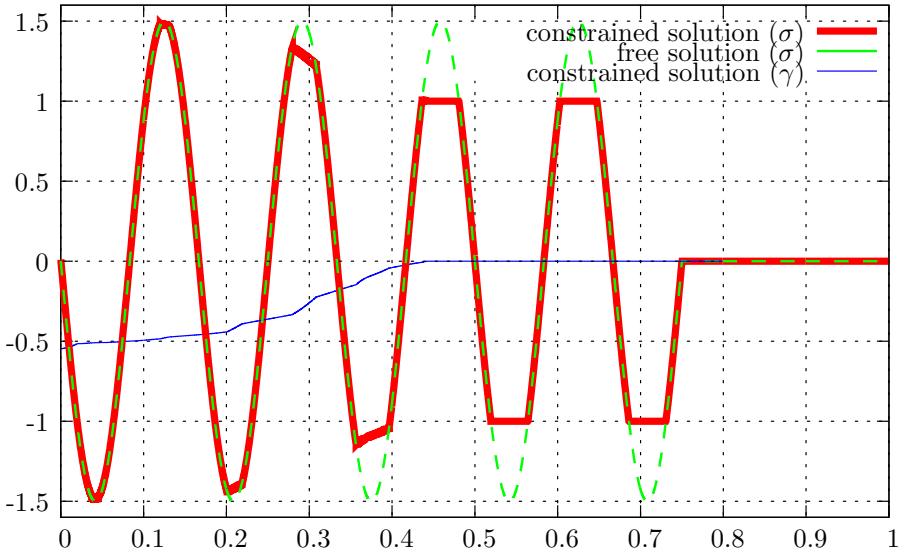

Figure 5: Solutions at time $t = 0.75$.

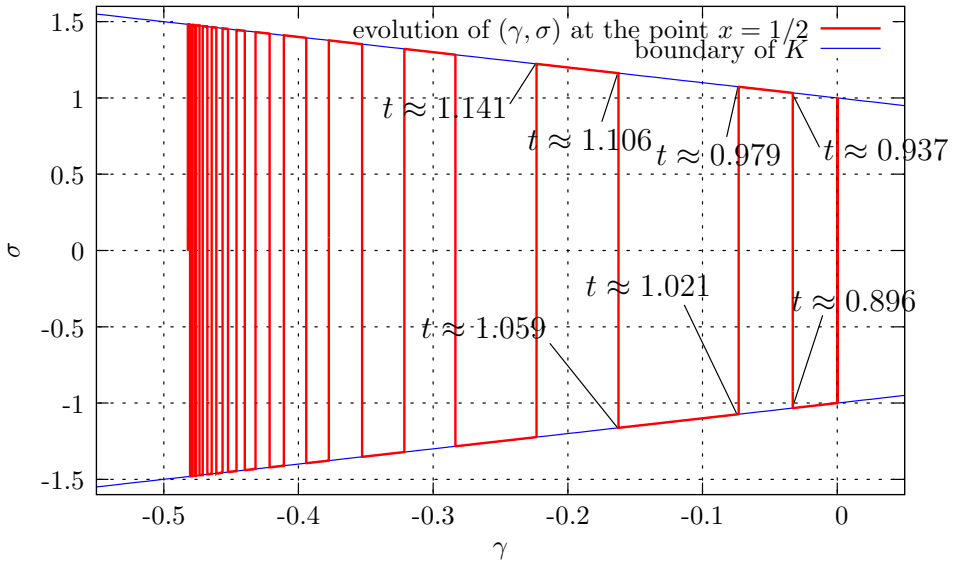

Figure 6: Parametric representation of the function $t \mapsto (\gamma(t, 1/2), \sigma(t, 1/2))$ for $0 \leq t \leq 3$.

where

$$\chi_\eta(t) = \begin{cases} 1 & \text{if } 0 \leq t < T, \\ (T-t)/\eta + 1 & \text{if } T \leq t < T + \eta, \\ 0 & \text{if } t \geq T + \eta, \end{cases}$$

and

$$\omega_\eta(t,x) = \begin{cases} 1 & \text{if } |x| \le r + L(T-t), \\ (r + L(T-t) - |x|)/\eta + 1 & \text{if } |x| \in r + L(T-t) + [0,\eta), \\ 0 & \text{if } |x| \ge r + L(T-t) + \eta. \end{cases}$$

Passing to the limit $\eta \to 0$ and since

$$\langle u - \widetilde{u}, A_i(u - \widetilde{u})\rangle \le L|u - \widetilde{u}|^2, \tag{59}$$

we obtain

$$-\int_{|x|<r} |u_\varepsilon(T,x) - \kappa|^2\,dx + \int_{|x|<r+LT} |u^0 - \kappa|^2\,dx \ge \frac{2}{\varepsilon}\int_A |P_K u_\varepsilon - u_\varepsilon|^2\,dx\,dt$$

where $A = \{(t,x) \in [0,T]\times\mathbb{R}, |x| \le r + L(T-t)\}$, which is included in $[0,T]\times B(0,r)$. Therefore, for all $\kappa \in K$,

$$\begin{aligned}\int_0^T \int_{|x|<r} |P_K u_\varepsilon - u_\varepsilon|^2\,dx\,dt &\le \frac{\varepsilon}{2}\int_{|x|<r+LT} |u^0 - \kappa|^2\,dx \\ &\le \frac{\varepsilon}{2}\int_{|x|<r+LT} |u^0 - \overline{u^0}|^2\,dx\end{aligned}$$

where

$$\overline{u^0} = \frac{1}{|B(0,r+LT)|}\int_{|x|<r+LT} u^0(x)dx$$

since $u^0 \in K$. To conclude, it suffices to use the Poincaré-Wirtinger inequality

$$\int_{|x|<r+LT} |u^0 - \overline{u^0}|^2\,dx \le C|u^0|^2_{H^1(B(0,r+LT))},$$

which finally provides (56). □

Now, assume that the solution u_ε of the relaxation problem (2) converges in L^2 to a function $\bar{u}$ when ε tends to 0 (using a priori H^1 bounds for instance). Then lemma 21 implies that $\bar{u} \in K$. Moreover, due to the properties of the projection P_K, u_ε satisfies inequalities (6) for all $\kappa \in K$ (see (58)) and so does $\bar{u}$. Therefore, $\bar{u}$ is the weak constrained solution.

References

[1] F. Berthelin and F. Bouchut, Weak solutions for a hyperbolic system with unilateral constraint and mass loss, Ann. Inst. H. Poincaré Anal. Non Linéaire 20 (2003),975–997.

[2] G. Boillat and T. Ruggeri, Hyperbolic principal subsystems: entropy convexity and subcharacteristic conditions. Arch. Rational Mech. Anal. 137 (1997), no. 4, 305–320.

[3] L. Brun, Introduction à la Thermodynamique des Matériaux, École Polytechnique, France, 1989.

[4] C. Cancès and T. Gallouët, On the time continuity of entropy solutions, to appear in J. Evolution Eq. (arXiv:0812.4765v2).

[5] G. Q. Chen, C. D. Levermore and T.-P. Liu, Hyperbolic conservation laws with stiff relaxation and entropy, Communications in Pure an Applied Mathematics, 787–830, 1994.

[6] C. M. Dafermos. The entropy rate admissibility criterion for solutions of hyperbolic conservation laws. J. Differential Equations 14 (1973), 202–212.

[7] G. Dal Maso, A. DeSimone, M. G. Mora and M. Morini, Arch. Ration. Mech. Anal., A vanishing viscosity approach to quasistatic evolution in plasticity with softening, Arch. Ration. Mech. Anal. 189 (2008), 469–544.

[8] B. Després, A Geometrical Approach to Nonconservative Shocks and Elastoplastic Shocks. Volume 186, Number 2, pp. 275-308, 2007.

[9] D. S. Drumheller, Introduction to wave propagation in nonlinear fluids and solids, Cambridge Press University, 1998.

[10] G. Duvaut and J.-L. Lions, Inequalities in mechanics and physics, Berlin; New York: Springer-Verlag 219, 1976.

[11] R. Eymard, T. Gallouët and R. Herbin, Finite volume methods, in Handbook of numerical analysis, Vol. VII, Handb. Numer. Anal., VII, North-Holland, Amsterdam, 2000, pp. 713–1020.

[12] E. Godlewski and P.-A. Raviart, *Hyperbolic systems of conservation laws*, Paris Ellipse, 1991.

[13] V. Jovanovic and C. Rohde, Finite-volume schemes for Friedrichs systems in multiple space dimensions: a priori and a posteriori error estimates. Numer. Methods Partial Differential Equations 21 (2005), no. 1, pp. 104–131.

[14] S. N. Kruzhkov, First order quasilinear equations with several independent variables, Mat. Sb. (N.S.), 81 (1970), pp. 228–255.

[15] A. G. Kulikovski, N. V. Pogorelov and A. Yu. Semenov, Mathematical aspects of numerical solution of hyperbolic systems, Chapman and Hall, Monographs and surveys in pure and applied mathematics, 118, 2001.

[16] N. N. Kuznetsov, The accuracy of certain approximate methods for the computation of weak solutions of a first order quasilinear equation, Ž. Vyčisl. Mat. i Mat. Fiz. **16** (1976), no. 6, 1489–1502, 1627.

[17] P. D. Lax, Hyperbolic systems of conservation laws and the theory of shock waves, SIAM, Philadelphia, 1973.

[18] P. Le Floch, Shock waves for nonlinear hyperbolic systems in non conservative forms, IMA Preprint Series 593, 1989.

[19] P. Le Floch, Entropy weak solutions to nonlinear hyperbolic systems under non conservative form, Comm. In P.D.E., 13(6), 669-727, 1988.

[20] Dal Maso, Le Floch and F. Murat, Definition and weak stability of non conservative products, J. Math. Pures Appl., 74 (1995), 458-483.

[21] I. Müller and T. Ruggeri, Rational Extended Thermodynamics, Springer Tracts in Natural Philosophy, 37. Springer-Verlag, New York, 1998..

[22] R. I. Nigmatulin, Dynamics of multiphase media, HPC, Friedly Editor, 1991.

[23] A. Nouri and M. Rascle, A global existence and uniqueness theorem for a model problem in dynamic elasto-platicity with istropic strain-hardening, Siam J. Math. anal., 26, 4, 850-868, 1995.

[24] V. M. Sadovskii, The theory of shock waves in compressible plastic media, Mechanics of solids A. 2001, vol. 36, no. 5, pp. 67-74.

[25] D. Serre, Systems of conservation laws I and II, Translated from the 1996 French original by I. N. Sneddon. Cambridge University Press, Cambridge, 1999.

[26] J. Simon, Compact sets in the space $L^p(0,T;B)$, Ann. Mat. Pura Appl. (4), 65–96 (1987).

[27] P. Suquet, Sur les équations de la plasticité : existence et régularité des solutions, J. Mécanique 20 (1981), 3–39.

[28] P. Suquet, Un espace fonctionnel pour les équations de la plasticité, Annales de la faculté des sciences de Toulouse Sér. 5, 1 no. 1 (1979), p. 77-87.

[29] J.-P. Vila and P. Villedieu, Convergence of an explicit finite volume scheme for first order symmetric systems. Numer. Math. 94 (2003), no. 3, pp. 573–602.