Controlling the spurious oscillations in a least squares formulation of the transport equation approximated with space-time finite element

Khadidja Benmansour[1], Elie Bretin[2], Loic Piffet[3], Jérôme Pousin[4]

[1]Université Abou Bekr Belkaid de Tlemcen, Université de Tlemcen BP 119 Algérie, jija_tlm@yahoo.fr
[2,3,4]Université de Lyon UMR-CNRS 5208 Institut Camille Jordan, 20 av. A. Einstein F-69100 Villeurbanne, elie.bretin@insalyon.fr, loic.piffet@yahoo.fr, jerome.pousin@insa-lyon.fr

ABSTRACT. Finite element methods are known to produce spurious oscillations when the transport equation is solved. In this paper, a variational formulation for the transport equation is proposed, and by introducing extremal values constraints combined with a penalization of the total variation of the solution, the spurious oscillations are cancelled for a first order Lagrange finite element method.

KEYWORDS. transport equation; space-time formulation; least squares; total variation; finite element; numerical oscillations; constrained optimization; image processing

1 Introduction

This paper concerns the control of spurious oscillations which take place when the transport equation is numerically solved with a space-time integrated least squares formulation in a finite element context. The transport equation (called optical flow equation in imaging modeling), subjected to some constraints, is widely used in imaging processes, see [5], [16] for example, or in computational anatomy [30]. A space-time integrated least squares formulation is well adapted in such a context since the problem is formulated as an optimization problem which allows to account for a large variety of constraints. Moreover, this formulation is known not to add diffusion in the orthogonal directions of integral curves, which is crucial when transporting functions with abrupt variations. The least squares method has been studied with Galerkin discontinuous finite element for the transport equation in [29] and is also widely used for solving partial differential equations, see [18] or [20] for elasticity and fluid mechanics problems, and finally [13] for some applications in a finite element context.

The maximum principle for PDE is known to be sufficient to guarantee positivity, monotonicity, and to not increasing the total variation of the solution. The space-time least squares formulation for the transport equation in an appropriate functional space satisfies a weak maximum principle (see theorem 2.4). On the contrary, an important feature of the finite element method for simulating transport phenomena is its inability to satisfy the weak maximum principle on general meshes for the standard Galerkin formulation. This deficiency manifests itself in spurious oscillations, which is a well known phenomenon in fluid dynamics.

Many remedies are available for finite difference techniques such as: the slope limiter and the flux corrector, which have been recently extended to the finite element method in [24]. The time partial differential operator of the transport equation is isolated in order to take advantage of the flow proper-

This work was supported by the LABEX MILYON (ANR-10-LABX-0070) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR); and by ANR 11-TecSan002-03.

ties of a differential equation, and the partial differential operator in space is treated in a specific way. Unfortunately, this technique does not extend to the formulation considered here, a space-time least-squares formulation, because time and space are handled in the same way.

The key idea for recovering a discrete maximum principle and for canceling spurious oscillations that we propose in this paper, is to transform the problem to an inequality constrained optimization problem, in order to handle the total variation and the extremal values of the solution. Combining a penalization of the total variation, where the value of the penalization parameter is moderate, with an extreme values constraint for the solution to the transport equation allows to cancel the spurious oscillations and not to change to much the shape of the solution.

Let us quote some recent works which combine a discretization method with optimization to enforce a maximum principle or some physical features of the solution: [14], [17], [26]. Another approaches such as the L^p minimization of the residual of the transport equation [21], or the L^1 minimization [22], or the iteratively reweighed least-squares [23] satisfy the maximum principle at a discret level, but, are more appropriate for discontinuous solutions to the transport equation.

Our work belongs to a recent stream of works that combine a discretization method with optimization to enforce valuable property of the solution. Because we aim to use our method for image processing, and since the TV norm is relevant in that context, we chose to add the TV norm of the solution to the L2 residual of the transport equation. In what follows, some properties concerning the recursive median filter are recalled, in order to have a better understanding of the role of the TV penalization parameter in the method we propose.

Minimization of the total variation (TV) under L^1 data fidelity, which consists of finding u solving

$$\underset{u}{\text{Argmin}} \quad \|u - f\|_{L^1} + k \|Du\|_{L^1} \qquad [1.1]$$

for a fixed k and a given f, has been wildly studied. It is well-known that solutions to that problem belong to the space of bounded variations functions. In [1] the study of problem (1.1) is restricted to the one dimensional case and to the discrete case. One solution to the minimization problem (1.1) discretized with a finite differences method can be obtained with a recursive median filter (RMF) of length 2k+1, which rectifies a regular signal f (i.e. a step function) perturbed with an overshoot (see Theorem 7 in [1]). An important feature of the RMF is the following: if the support of the perturbation is less than k, the initial signal is recovered, if the support of the perturbation is more than k, the signal is partially rectified, and its shape is modified. For the oscillations of functions the situation is more intricate, nevertheless, there are damped with the RMF of a sufficient length [1] . It is worth to notice that, the more the length of the RMF is large the more the shape of the function is changed. Owing to these results, we aim to distinguish the overshoots attached to the extremal values of the transport equation solution (these extreme values are given by the boundary condition as a consequence of the weak maximum principle) with the other existing oscillations. The overshoots will be cancelled by using a projection method because they can have a large support, and the existing oscillations will be controlled through the total variation of the solution to the transport equation.

In section 2 a description of the problem is given and function spaces are introduced. Then a weak maximum principle is recalled (see theorem 2.4). In section 3, the hypothesis that Ω is a rectangle is assumed. This hypothesis is not too restrictive in the context of imaging. A Galerkin formulation with Lagrange finite element $\mathbb{Q}^1$ is introduced, and some numerical experiments are presented for a time

stepping strategy. Unfortunately the computed solution exhibits spurious oscillations, therefore a basic strategy which consists in restricting the solution to a convex set in order to cancel the perturbations attached to the extremal values is presented. The end of the section is dedicated to the implementation of a projection technique on the set of nonnegative functions, the convergence of which is analyzed. Finally in section 4 a formulation, which penalizes the total variation of the solution and ensures its non negativity, is introduced and an existence result is given. A FISTA algorithm is choosed for computing the solution, and numerical experimentations in 1D or 2D are presented, demonstrating that the spurious oscillations present in the previous numerical schemes do not exist anymore.

2 The problem description and the functional setting

Let $\Omega \subset \mathbb{R}^d$ (with $d = 1; 2; 3$) be a bounded domain with a Lipschitz boundary $\partial\Omega$ satisfying the cone property. Throughout this article, the Euclidean inner product will be denoted by $(\,.\,|\,.\,)$. If $T > 0$ is given, set $Q = \Omega\times]0, T[$. Consider an advection velocity $v : Q \to \mathbb{R}^d$ and $f : Q \to \mathbb{R}$ a given source term. In all of the paper, the velocity v has at least the following regularity

$$v \in L^\infty(Q)^d \ \text{ and } \ \text{div}\,(v) \in L^\infty(Q). \qquad [2.1]$$

Let

$$\Gamma_- = \{x \in \partial\Omega : (\,v(x,t)\,|\,n(x)\,) < 0,\}$$

where $n(x)$ is the outer normal to $\partial\Omega$ at point x. For the sake of simplicity, it is assumed that Γ_- does not depend on t.

The problem consists in finding a function $c : Q \to \mathbb{R}$ satisfying the following partial differential equation

$$\partial_t c + (\,v(x,t)\,|\,\nabla c(x,t)\,) = f(t,x) \quad \text{in } Q, \qquad [2.2]$$

and the initial and inflow boundary conditions

$$c(x,0) = \ c_0(x) \qquad \text{for } x \text{ in } \Omega \qquad [2.3]$$

$$c(x,t) = \ c_1(x,t) \qquad \text{for } x \text{ on } \Gamma_-. \qquad [2.4]$$

As usual, when c_1, c_0 and v are sufficiently regular, changing the source term f if necessary, one can assume that $c_1 = 0$ on Γ_-, and $c_0 = 0$ on Ω. A similar result will be given later, using a suitable trace theorem. In what follows, the functional setting for a variational formulation of the problem (2.2–2.4) will be given, (see also [6, 7]). Moreover a trace operator is recalled in this context.

For $v \in L^\infty(Q)^d$, with $\text{div}\,(v) \in L^\infty(Q)$, define

$$\widetilde{v} = (1, v_1, v_2, \dots, v_d)^t \in L^\infty(Q)^{d+1}$$

and for a sufficiently regular function φ defined on Q, set

$$\widetilde{\nabla}\varphi = \left(\frac{\partial\varphi}{\partial t}, \frac{\partial\varphi}{\partial x_1}, \frac{\partial\varphi}{\partial x_2}, \dots, \frac{\partial\varphi}{\partial x_d}\right)^t,$$

and $\widetilde{n}$ denotes the outward unit vector on ∂Q. Let now

$$\begin{aligned} \partial Q_- &= \{(x,t) \in \partial Q, (\widetilde{u}\,|\,\widetilde{n}\,) < 0\} \\ &= \Gamma_- \times (0,T) \cup \Omega \times \{0\}, \end{aligned}$$

and set

$$c_b(x,t) = \begin{cases} c_0(x) & \text{if} \quad (x,t) \in \Omega \times \{0\} \\ c_1(t,x) & \text{if} \quad (x,t) \in \Gamma_- \times (0,T). \end{cases} \qquad [2.5]$$

Here it is assumed that $c_b \in L^2(\partial Q_-)$. For $\varphi \in \mathcal{D}(\overline{Q})$, the infinitely differentiable functions in $\overline{Q}$, consider the norm

$$\|\varphi\|_{H(v,Q)} = \left(\|\varphi\|^2_{L^2(Q)} + \left\| \left(\widetilde{v}\,|\,\widetilde{\nabla}\varphi \right) \right\|^2_{L^2(Q)} + \int\limits_{\partial Q_-} |\,(\,\widetilde{v}\,|\,\widetilde{n}\,)\,|\varphi^2\, d\widetilde{\sigma} \right)^{1/2},$$

(see also [6, 7, 8, 11]) and then define the space $H(v,Q)$ as the closure of $\mathcal{D}(\overline{Q})$ for this norm:

$$H(v,Q) = \overline{\mathcal{D}(\overline{Q})}^{H(v,Q)}$$

If v is regular enough, it can be seen that

$$H(v,Q) = \left\{ \rho \in L^2(Q), \left(\widetilde{v}\,|\,\widetilde{\nabla}\rho \right) \in L^2(Q), \rho|_{\partial Q_-} \in L^2(\partial Q_-, |\,(\,\widetilde{v}\,|\,\widetilde{n}\,)\,|\,d\widetilde{\sigma}) \right\}$$

(see e.g. [25, 19]). In proposition 2 from [12] a trace operator $\gamma_{\widetilde{n}_-}$ with values in $L^2(\partial Q_-, |\,(\,\widetilde{v}\,|\,\widetilde{n}\,)\,|\,d\widetilde{\sigma})$ is defined for function belonging to $H(u,Q)$, allowing to introduce the following space:

$$\begin{aligned} H_0 = H_0(u,Q,\partial Q_-) &= \{\rho \in H(u,Q),\ \rho = 0 \text{ on } \partial Q_-\} \\ &= H(u,Q) \cap \text{ Ker } \gamma_{\widetilde{n}_-}. \end{aligned}$$

We now recall an extension of the *curved Poincaré inequality* obtained in [6, 7].

THEOREM 2.1.– *If $v \in L^\infty(Q)^d$ and div $(v) \in L^\infty(Q)$, the semi-norm on $H(u,Q)$ defined by*

$$|\rho|_{1,v} = \left(\int\limits_Q \left(\widetilde{v}\,|\,\widetilde{\nabla}\rho \right)^2 dx\, dt + \int\limits_{\partial Q_-} |\,(\,\widetilde{v}\,|\,\widetilde{n}\,)\,|\rho^2\, d\widetilde{\sigma} \right)^{1/2} \qquad [2.6]$$

is a norm, equivalent to the norm given on $H(v,Q)$.

Henceforth the space $H(v,Q)$ is equipped with the norm $|\varphi|_{1,v}$.

REMARK 2.2.– *a) Using the above result, if $c_b = 0$, the semi-norm*

$$|\rho|_{1,v} = \left(\int\limits_Q \left(\widetilde{v}\,|\,\widetilde{\nabla}\rho \right)^2 dx\, dt \right)^{1/2}$$

is a norm on H_0 which is equivalent to the usual norm on $H(v,Q)$.

Let us end this section with a least squares formulation in $L^2(Q)$. A space-time least squares solution of equation (2.2) corresponds to a minimizer in
$\{\varphi \in H(v,Q); \gamma_{\widetilde{n}_-}(\varphi) - c_b = 0\}$ of the following convex, $H(v,Q)$-coercive functional

$$J(c) = \frac{1}{2}\left(\int_Q \left(\left(\widetilde{v} \,|\, \widetilde{\nabla} c \right) - f \right)^2 dx\, dt - \int_{\partial Q_-} c^2 \left(\widetilde{u} \,|\, \widetilde{n} \right) d\widetilde{\sigma} \right).$$

The Gâteau derivative of J is

$$DJ(c)\varphi = \int_Q \left(\left(\widetilde{v} \,|\, \widetilde{\nabla} c \right) - f \right) \left(\widetilde{v} \,|\, \widetilde{\nabla} \varphi \right) dx\, dt - \int_{\partial Q_-} c\varphi \left(\widetilde{v} \,|\, \widetilde{n} \right) d\widetilde{\sigma}.$$

So a sufficient condition to get the least squares solution of (2.2–2.4) is the following *weak formulation*: If $c_b \in L^2(\partial Q_-)$, find $c \in H(v,Q)$ such that

$$\begin{aligned} \int_Q \left(\widetilde{v} \,|\, \widetilde{\nabla} c \right) \left(\widetilde{v} \,|\, \widetilde{\nabla} \varphi \right) dx\, dt &= \int_Q f \left(\widetilde{v} \,|\, \widetilde{\nabla} \varphi \right) dx\, dt, \\ \gamma_{\widetilde{n}_-}(c) &= c_b, \end{aligned} \qquad [2.7]$$

for all $\varphi \in H_0$ (see [6, 7, 8, 11, 27]).

2.1 *A weak maximum principle for the space-time least squares formulation*

In this subsection a weak maximum principle is given. First, let us define a penalized formulation of the space-time least squares, useful for some L^∞ estimates.

LEMMA 2.3.– *If $c_b \in L^2(\partial Q_-)$, let c^m be the solution of*

$$\int_Q \left(\widetilde{v} \,|\, \widetilde{\nabla} c^m \right) \left(\widetilde{v} \,|\, \widetilde{\nabla} \varphi \right) dx\, dt - m \int_{\partial Q_-} (c^m - c_b)\varphi \left(\widetilde{v} \,|\, \widetilde{n} \right) d\widetilde{\sigma} = \int_Q f \left(\widetilde{v} \,|\, \widetilde{\nabla} \varphi \right) dx\, dt, \qquad [2.8]$$

$\forall \varphi \in H(v,Q)$. There is a subsequence of c^m which converges weakly in $H(v,Q)$ to the solution c of (2.7) when m goes to infinity.

A weak maximum principle is proved for the penalized formulation given in the lemma (2.3) and is extended to the problem (2.7) solution (see [12] p. 167).

THEOREM 2.4.– *Assume that the function $f = 0$ in equation (2.7) and that the function $c_b \in L^\infty(\partial Q_-)$. Then the solution of equation (2.7) satisfies*

$$\inf c_b \leq c \leq \sup c_b.$$

3 Finite element approximation

From now on it is assumed that Ω is a rectangle. This hypothesis is not too restrictive in the context of imaging.

Let $\{\varphi_1, \varphi_2, \dots, \varphi_N\}$ be a basis of the finite element subspace $V_h \subset H_0$, obtained, for example,with a rectangular structured mesh $\mathcal{T}_h$ of the domain Q, with first order $\mathbb{Q}_1$ Lagrange quadrilateral finite element:

$$Q = \bigcup_{\mathcal{T}_h} T, \quad V_h = \left\{ \varphi \in C^0(\overline{Q}) \mid \varphi_{|T} \in \mathbb{Q}^1(T) \right\}. \tag{3.1}$$

Define the bilinear symmetric form $a(\cdot,\cdot)\, V_h \times V_h \to \mathbb{R}$ by

$$a(\psi_h, \varphi_h) = \int_Q \left(\widetilde{v} \,|\, \widetilde{\nabla}\psi_h \right) \left(\widetilde{v} \,|\, \widetilde{\nabla}\varphi_h \right) dxdt.$$

For handling the boundary condition since the regular function c_b admits an extension C_b in Q belonging at least to $H(v, Q)$, the right hand side is changed in $f - \left(\widetilde{v} \,|\, \widetilde{\nabla} C_b \right)$. An approximation of the problem (2.7), consists in finding $c_h \in V_h$ such that

$$a(\varphi_h, c_h) \;=\; \int_Q \left(f - \left(\widetilde{v} \,|\, \widetilde{\nabla} C_b \right) \right) \left(\widetilde{v} \,|\, \widetilde{\nabla}\varphi_h \right) dx\, dt; \tag{3.2}$$

for all $\varphi_h \in V_h$, where $c_h = \sum_{j=1}^{N} \varphi_j(t,x) \cdot c_j$. With these notations, equation (3.2) becomes

$$\sum_{j=1}^{N} c_j \int_Q \left(\widetilde{v} \,|\, \widetilde{\nabla}\varphi_j \right) \left(\widetilde{v} \,|\, \widetilde{\nabla}\varphi_i \right) dx\, dt = \int_Q \left(f - \left(\widetilde{v} \,|\, \widetilde{\nabla} C_b \right) \right) \left(\widetilde{v} \,|\, \widetilde{\nabla}\varphi_i \right) dx\, dt; \tag{3.3}$$

for all $i = 1, \dots, N$. Set for all $1 \le i, j \le N$

$$a_{ij} = \int_Q \left(\widetilde{v} \,|\, \widetilde{\nabla}\varphi_j \right) \left(\widetilde{v} \,|\, \widetilde{\nabla}\varphi_i \right) dx\, dt; \quad b_i = \int_Q \left(f - \left(\widetilde{v} \,|\, \widetilde{\nabla} C_b \right) \right) \left(\widetilde{v} \,|\, \widetilde{\nabla}\varphi_i \right) dx\, dt.$$

The coefficients, a_{ij}, b_i, are computed in the standard way. If $A = (a_{ij})_{1 \le i,j \le N}$, $B = (b_i)_{1 \le i \le N}$ and $C = (c_i)_{1 \le i \le N}$, then the solution of the linear system

$$AC = B \tag{3.4}$$

is the solution of problem (3.2). The method we consider is a marching technique, that is to say the solution is computed time slice by time slice. Since the first component of the velocity $\widetilde{v}_1 = 1$, such a strategy is possible because the integral curves associated to $\widetilde{v}$ are increasing with respect to time. At each time step we solve a 'local time' problem where the initial condition is c_h at the current time step and the unknown is c_h at the next time step. The stiffness matrix of the system is calculated for a single slice of finite elements of width $\triangle t$. Here, the solution c_h comprising only the solutions computed at time t (taken as an initial condition or boundary condition) and at time $t + \triangle t$. The system is solved and we proceed step by step until the final time is reached. Assume the domain $Q = \Omega \times \triangle t$, as mentioned above, and let $V_h \subset H(v, Q)$ be a first order $\mathbb{Q}^1$ Lagrange finite element subspace. The vector C is

decomposed as follows: C_- containing only values at nodes at time t and the unknowns at time $t + \triangle t$ are stored in $\overline{C}$, such as degrees of freedom belonging to ∂Q_- have the lower labels. We have:

$$AC = \begin{pmatrix} M & N \\ P & Q \end{pmatrix} \begin{pmatrix} C_- \\ \overline{C} \end{pmatrix} = \begin{pmatrix} B_- \\ \overline{B} \end{pmatrix} \quad [3.5]$$

The solution $\overline{C}$ at time step $t + \triangle t$ is obtained by solving the reduced system:

$$Q\overline{C} = \overline{B} - PC_-. \quad [3.6]$$

3.1 *Numerical experiments for the 1D time marching scheme*

In this section, a numerical example with the transport equation with $v = 1$ and with only a non zero initial condition, taking the initial condition as a step of height 1 located at $x = 0.5$ and given by $c_0(x,t) = \frac{1}{2}(1 - tanh(100x - 50))$, is investigated. For that case, the solution to the transport equation is a step which translates from left to right without deformation as time evolves. In order to evaluate the numerical scheme properties, the computed solution is plotted for different times. The numerical scheme analyzed here is a convergent scheme, that is to say, the $H(v,Q)$-norm, introduced in section 2, of the error between the solution to the transport equation and the computed one, goes to zero when the discretization parameter in space and in time (the size of the mesh $\frac{1}{n} \times \triangle t$) goes to zero.

The number of elements is $n = 80$ and the time step is $\triangle t = 1/79$. In the following figure 4.3, the horizontal axis is the x axis, and the values of the computed solution are reported on the vertical axis. The computed solution is represented after 1 and 20 time steps.

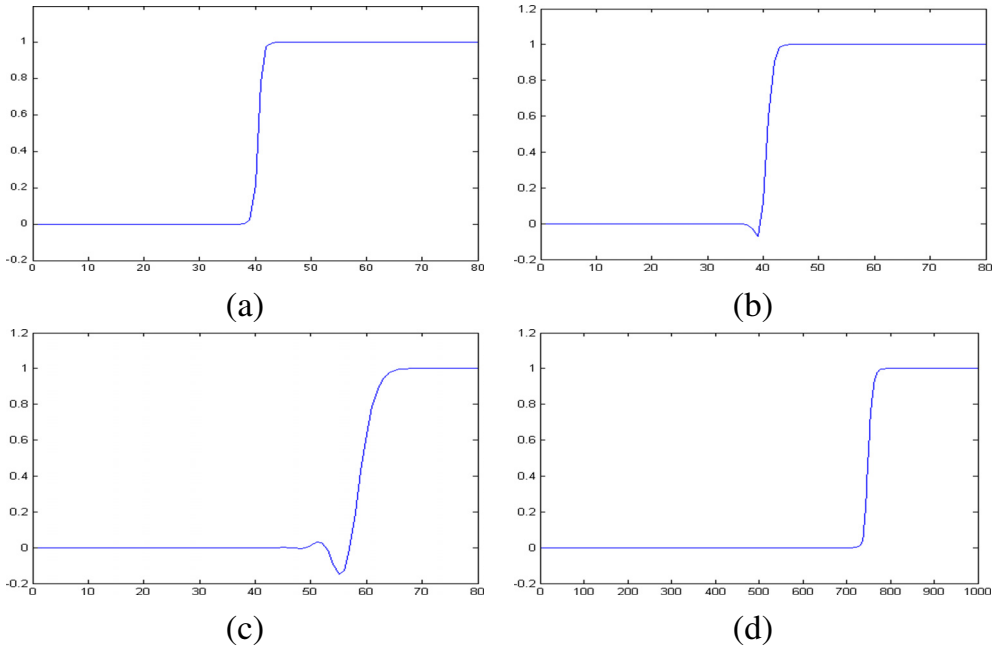

Figure 3.1: *(a) Initial signal; (b) Solution for one time step; (c) Solution for 20 time steps; (d) convergence of the solution for n=1000.*

In figure 4.3 the solution exhibits undershoots as time elapses. Despite these oscillations, the method converges when n goes to infinity see 4.3-d where the computed solution is plotted at the same time as in 4.3-c, but for much smaller discretization parameters. The discrete maximum principle is strongly

related to the M-matrix property and to the sign of the right hand side (r.h.s. in short) of the system that is to be solved. In (3.5) since the velocity is constant the M-matrix property holds true (which will be no longer true for a variable velocity or in 2D case), but the sign property of the r.h.s is not satisfied. To summarize, the Lagrange finite element method, associated to space-time least squares formulations for the transport equation converges when the discretization parameters go to zero, but exhibits spurious oscillations.

3.2 *Projection method with a Generalized Lagrange multiplier method*

A least squares method associated with Lagrange finite element for computing the solution to the transport equation yields numerical solutions polluted by spurious perturbations. In this subsection, dealing with the simplified example considered previously, a projection onto the cone of nonnegative functions K_h is investigated as remedy to the perturbations around minimal value. The results presented below could be extended to more complex constraints. For example, if the baseline is not zero but a regular function, by translating the solution, the problem is reduced to a non negativity constraint. More complex convex subsets could also be handled. The idea to impose to the solution to belong to a convex subset requires, to have an efficient projection strategy. The main concern of this subsection is the implementation of a nodal projection strategy in the context of a least squares formulation. The nonnegativity of the basis functions φ_k, $1 \geq k \geq N$ allows us to define K_h as follows :

$$K_h = \left\{ \sum_{k=1}^{N} \alpha_k \varphi_k \;\middle|\; \alpha_k \in \mathbb{R}^+ \right\}. \qquad [3.7]$$

Let us rephrase the problem (3.4) as a constrained optimization problem in $\mathbb{R}^N$. Find C_p satisfying :

$$\begin{cases} AC_p = B \\ C_p \geq 0. \end{cases} \qquad [3.8]$$

Express the previous problem as the following minimization problem on $\mathbb{R}^N_+$:

$$C_p = \underset{X \in \mathbb{R}^N_+}{\text{Argmin}} \quad \frac{1}{2} X^t A X - X^t B. \qquad [3.9]$$

The problem is well posed, and by using the complementarity conditions [28]

$$0 \leq (AC_p - B); C_p \perp (AC_p - B) \qquad [3.10]$$

problem (3.9) can be be computed by using the following generalized Lagrange multiplier method for $0 < r$ fixed.

$$\begin{cases} AC_p = B + \Lambda \\ \Lambda = (\Lambda - rC_p)^+ \end{cases} \qquad [3.11]$$

where the positive part of a vector denotes the positive part of its components. An iterative algorithm is proposed for solving the problem (3.11). Set $C_p^0 = 0_{\mathbb{R}^N}; \Lambda^0 = 0_{\mathbb{R}^N}$, then compute:

$$\begin{cases} AC_p^{k+1} = B + \Lambda^{k+1} \\ \Lambda^{k+1} = (\Lambda^k - rC_p^k)^+ \end{cases} \qquad [3.12]$$

Now the convergence of the iterative procedure is proved. We have:

LEMMA 3.1.– *Let μ_1 be the first eigenvalue of the positive definite matrix A. For all r verifying $0 < r < 2\mu_1$ the algorithm 3.12 converges.*

Proof.

$$\left\{ \begin{array}{l} A(C_p^{k+1} - C_p^k) = \Lambda^{k+1} - \Lambda^k \\ \Lambda^{k+1} - \Lambda^k = (\Lambda^k - rC_p^k)^+ - (\Lambda^{k-1} - rC_p^{k-1})^+. \end{array} \right. \quad [3.13]$$

Since z^+ is a 1-lipschitzian function we deduce:

$$\left\|\Lambda^{k+1} - \Lambda^k\right\|^2 \leq \left\|\Lambda^k - \Lambda^{k-1}\right\|^2 - 2r\left(\Lambda^k - \Lambda^{k-1} \,|\, C_p^k - C_p^{k-1}\right) + r^2\left\|C_p^k - C_p^{k-1}\right\|^2.$$

Since the matrix A is positive definite, by using the first equation of (3.13), the previous inequality becomes:

$$\begin{aligned} \left\|\Lambda^{k+1} - \Lambda^k\right\|^2 &\leq \left\|\Lambda^k - \Lambda^{k-1}\right\|^2 - 2r\left(A(C_p^k - C_p^{k-1}) \,|\, C_p^k - C_p^{k-1}\right) + r^2\left\|C_p^k - C_p^{k-1}\right\|^2 \\ &\leq \left\|\Lambda^k - \Lambda^{k-1}\right\|^2 + r(r - 2\mu_1(A))\left\|C_p^k - C_p^{k-1}\right\|^2. \end{aligned}$$

If $\left\|C_p^k - C_p^{k-1}\right\|^2 = 0$, the sequence $\{\Lambda^p\}_{p>k}$ becomes stationary and thus converges. If $\left\|C_p^k - C_p^{k-1}\right\|^2 \neq 0$ we have:

$$\left\|\Lambda^{k+1} - \Lambda^k\right\|^2 < \left\|\Lambda^k - \Lambda^{k-1}\right\|^2,$$

thus there exists $0 < \xi < 1$ such that:

$$\left\|\Lambda^{k+1} - \Lambda^k\right\|^2 \leq \xi\left\|\Lambda^k - \Lambda^{k-1}\right\|^2.$$

For all $q < p$ we deduce:

$$\|\Lambda^p - \Lambda^q\|^2 \leq \sum_{l=q+1}^{l=p}\left\|\Lambda^l - \Lambda^{l-1}\right\|^2 \leq \sum_{l=q+1}^{l=p}\xi^{l-1}\left\|\Lambda^1\right\|^2 \leq \xi^q\sum_{m=0}^{\infty}\xi^m\left\|\Lambda^1\right\|^2$$

which proves that $\{\Lambda^k\}_{k\in\mathbb{N}}$ is a Cauchy sequence.

The sequence $\{C_p^k\}_{k\in\mathbb{N}}$ is also a Cauchy sequence, and we can take the limit in the equations.

□

Considering as before the 1D example of the moving step from left to right, the obtained numerical results are presented in figure 3.2 in the first row for one and for 20 time steps. The undershoot is cancelled.

If we investigate the example of a double step, the height of which is 2, moving from left to right, the generalized Lagrange multiplier method is not able to handle spurious oscillations as it is shown on the second row of figure 3.2 where the following initial condition: $c_0(x,t) = 1/2 + (1/2\tanh(100x - 20)) + 1/2 + (1/2\tanh(100x - 35))$ has been used. The computed solution with the generalized Lagrange multiplier method is presented after one time step and after 10 time steps with the same size of the mesh as before. This simple example is interesting because the function exhibits abrupt variations in its increasing and decreasing parts.

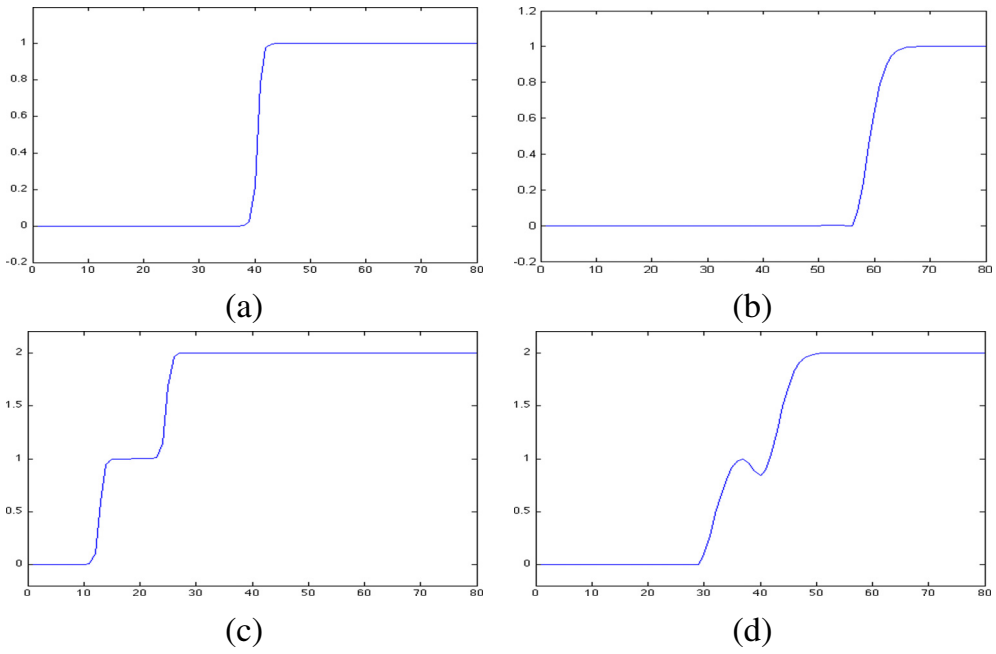

Figure 3.2: *First column : Initial signal; Second column : Time marching solution to generalized Lagrange multiplier (3.11) for 20 time steps for the first row, and for 10 time steps for the second row,*

4 A least squares formulation with penalized total variation and non negativity

In [1] it is checked that RMF is also efficient for controlling oscillations with a reduced change for the shape of the function, and that penalizing the total variation of the function acts in the same way. Oscillations are consequences of abrupt changes of sign and values of the gradient of the function to be transported. In this section the strategy of penalizing the TV and the non negativity of the computed solution is investigated for damping the spurious oscillations and the undershoot around the minimal value.

First, let us recall the main properties of the space of bounded variation functions $BV(Q)$ (see [2, 3, 4] for example). Introduce

$$TV(u) = \sup \left\{ \int_Q u(x) \operatorname{div} \xi(x)\, dx \mid \xi \in \mathcal{C}_c^1(Q),\ \|\xi\|_\infty \leq 1 \right\} \qquad [4.1]$$

and define the space

$$BV(Q) = \{u \in L^1(Q) \mid TV(u) < +\infty\}.$$

The space $BV(Q)$, endowed with the norm $\|u\|_{BV(Q)} = \|u\|_{L^1} + TV(u)$, is a Banach space. The derivative in distribution sense of a function $u \in BV(Q)$ is a bounded Radon measure, denoted Du, and $TV(u) = \int_Q |Du|$ is the total variation of u. Note that, for $u \in W^{1,1}(Q)$, $TV(u) = \|\nabla u\|_{L^1(Q)}$. We refer the reader to [2, 3] for more details and results.

Introduce K as the cone of nonnegative functions:

$$K = \{\varphi \in H_0 \cap BV(Q),\ \varphi \geq 0 \text{ a.e.}\}, \qquad [4.2]$$

and denotes by I_K its associated indicator function:

$$I_K(\varphi) = \begin{cases} 0 & \text{if } \varphi \in K \\ +\infty & \text{otherwise.} \end{cases}$$

For a given $\lambda \in \mathbb{R}_+$, we consider the following optimization problem:

$$u_\lambda = \underset{c \in H_0 \cap BV(Q)}{\operatorname{argmin}} \; J(c) + \lambda TV(c) + I_K(c) = \underset{c \in H_0 \cap BV(Q)}{\operatorname{argmin}} \; F(c), \qquad [4.3]$$

where the function J has been introduced in the first section

$$J(c) = \tfrac{1}{2} \int_Q \left(\left(\widetilde{v}(x,t) \,|\, \widetilde{\nabla} c(x,t) \right) - f(t,x) \right)^2 \, dx \, dt.$$

The following theorem gives an existence and uniqueness result for problem (4.3).

THEOREM 4.1.– *For all nonnegative λ, the problem 4.3 has a unique solution.*

Proof. Let $(c_n)_n \in H_0 \cap BV(Q)$ be a minimizing sequence of F, i.e.

$$\lim_{n \to +\infty} F(c_n) = \inf \{F(c) \mid c \in H_0 \cap BV(Q)\} < +\infty.$$

The sequence $(|c_n|_{1,v})_{n \in \mathbb{N}}$ is bounded, therefore the sequence $(c_n)_{n \in \mathbb{N}}$ converges weakly to $c^* \in H_0$, up to a subsequence. The function J is convex and lower semi continuous (l.s.c. in short), we have:

$$J(c^*) \leq \liminf_{n \to +\infty} J(c_n). \qquad [4.4]$$

The semi norm $|\cdot|_{1,v}$ and the norm $\|\cdot\|_{H(v,Q)}$ are equivalent in H_0, the sequence $(c_n)_{n \in \mathbb{N}}$ is bounded in $L^2(Q)$, and $(c_n)_{n \in \mathbb{N}}$ converges weakly towards c^* in $L^2(Q)$. Furthermore, Q is bounded, the embedding of $L^2(Q)$ into $L^1(Q)$ is continuous, $(c_n)_{n \in \mathbb{N}}$ is bounded in $L^1(Q)$ as well, and therefore it is bounded in $BV(Q)$. The compact embedding of $BV(Q)$ into $L^1(Q)$ yields the strong convergence in $L^1(Q)$ of a subsequence of $(c_n)_{n \in \mathbb{N}}$ towards c^*, with $c^* \in BV(Q)$.

From l.s.c. results of the TV operator, we have:

$$TV(c^*) \leq \liminf_{n \to +\infty} TV(c_n). \qquad [4.5]$$

The subspace K is convex and closed for the L^1-norm, I_K is thus convex and l.s.c. We conclude

$$I_K(c^*) \leq \liminf_{n \to +\infty} I_K(c_n). \qquad [4.6]$$

Finally we have up to a subsequence:

$$F(c^*) \leq \liminf_{n \to +\infty} F(c_n), \qquad [4.7]$$

that is to say

$$F(c^*) \leq \inf \{F(c) \mid c \in H_0 \cap BV(Q)\}. \qquad [4.8]$$

Uniqueness is a consequnece of the convexity.

□

Now let us give a technical result concerning the behavior of ρ_λ with respect to the parameter λ which will allow to impose an upper bound for the total variation.

LEMMA 4.2.– *Let ρ_λ be the solution to the problem 4.3. Then the mapping $Y : \lambda \mapsto \rho_\lambda$ from $\mathbb{R}_+$ with values in $H_0 \cap BV(Q)$ is continuous. Moreover, the application $T : \lambda \mapsto TV(\rho_\lambda)$ from $\mathbb{R}_+$ with values in $\mathbb{R}_+$ is continuous and decreasing towards zero.*

Proof. We first show that T is a decreasing function that converges to zero. Let $\lambda_2 > \lambda_1 > 0$ be given. We set $\rho_{\lambda_1} = Y(\lambda_1)$ and $\rho_{\lambda_2} = Y(\lambda_2)$. Since ρ_λ is the minimal argument of the functional F, we can write the two following inequalities :

$$J(\rho_{\lambda_1}) + \lambda_1 TV(\rho_{\lambda_1}) + I_K(\rho_{\lambda_1}) \leq J(\rho_{\lambda_2}) + \lambda_1 TV(\rho_{\lambda_2}) + I_K(\rho_{\lambda_2}), \quad [4.9]$$

and

$$J(\rho_{\lambda_2}) + \lambda_2 TV(\rho_{\lambda_2}) + I_K(\rho_{\lambda_2}) \leq J(\rho_{\lambda_1}) + \lambda_2 TV(\rho_{\lambda_1}) + I_K(\rho_{\lambda_1}). \quad [4.10]$$

The sum of these two inequalities gives

$$\lambda_1 TV(\rho_{\lambda_1}) + \lambda_2 TV(\rho_{\lambda_2}) \leq \lambda_1 TV(\rho_{\lambda_2}) + \lambda_2 TV(\rho_{\lambda_1}), \quad [4.11]$$

then

$$(\lambda_2 - \lambda_1)(TV(\rho_{\lambda_2}) - TV(\rho_{\lambda_1})) \leq 0, \quad [4.12]$$

that is to say

$$TV(\rho_{\lambda_2}) - TV(\rho_{\lambda_1}) \leq 0. \quad [4.13]$$

The decreasing of T follows.
Let $\lambda > 0$ be given. If we set $\rho_\lambda = Y(\lambda)$, we have, for all $\rho \in H_0 \cap BV(Q)$

$$J(\rho_\lambda) + \lambda TV(\rho_\lambda) + I_K(\rho_\lambda) \leq J(\rho) + \lambda TV(\rho) + I_K(\rho). \quad [4.14]$$

Specifying the previous inequality for $\rho = 0$ gives:

$$J(\rho_\lambda) + \lambda TV(\rho_\lambda) + I_K(\rho_\lambda) \leq J(0). \quad [4.15]$$

Sine J and I_K are nonnegative functions, we have

$$\lambda TV(\rho_\lambda) \leq J(0). \quad [4.16]$$

If $J(0) = 0$, then $TV(\rho_\lambda) = 0$ for all $\lambda > 0$. If $J(0) > 0$, $TV(\rho_\lambda) \leq \dfrac{J(0)}{\lambda}$. In both cases, $TV(\rho_\lambda) \underset{\lambda \to +\infty}{\longrightarrow} 0$, that is to say $T(\lambda) \underset{\lambda \to +\infty}{\longrightarrow} 0$.

Now we investigate the continuity of T. Let $\lambda \in \mathbb{R}_+^*$ and $(\lambda_n)_{n \in \mathbb{N}}$ be a sequence of elements of $\mathbb{R}_+$ that converges to λ. In what follows, it is proved that ρ_{λ_n} converges weakly toward ρ_λ. From the definition of ρ_{λ_n}, we have, for all $\rho \in H_0 \cap BV(Q)$ and $n > 0$,

$$J(\rho_{\lambda_n}) + \lambda_n TV(\rho_{\lambda_n}) + I_K(\rho_{\lambda_n}) \leq J(\rho) + \lambda_n TV(\rho) + I_K(\rho). \quad [4.17]$$

The sequence $(J(\rho)+\lambda_n TV(\rho))_{n\in\mathbb{N}}$ converges to $(J(\rho)+\lambda TV(\rho))$, so the sequence $(J(\rho_n))_n$ is bonded, and then $(\|\rho_{\lambda_n}\|_{H(v,Q)})_{n\in\mathbb{N}}$ is bounded as well. This implies that it exists $\rho^* \in H(v,Q)$ such that $(\rho_{\lambda_n})_{n\in\mathbb{N}}$ converges weakly in $H(v,Q)$ and therefore in $L^2(Q)$ to ρ^*. The sequence $(\rho_{\lambda_n})_{n\in\mathbb{N}}$ is bounded in $L^2(Q)$ and then in $L^1(Q)$, since Q is bounded. Moreover, $(TV(\rho_{\lambda_n}))_{n\in\mathbb{N}}$ is also bounded. Consequently, $(\rho_{\lambda_n})_{n\in\mathbb{N}}$ is bounded in $BV(Q)$. Accounting for compact embedding results, we get that $(\rho_{\lambda_n})_{n\in\mathbb{N}}$ strongly converges to $\rho^* \in BV(Q)$ in $L^1(Q)$, up to a subsequence.

Since $\rho \mapsto \left\| \left(\widetilde{v} \,|\, \widetilde{\nabla}\rho \right) - f \right\|^2_{L^2(Q)}$ is convex and l.s.c., we have

$$J(\rho^*) \leq \liminf_{n\to+\infty} J(\rho_{\lambda_n}), \tag{4.18}$$

and the l.s.c. property of the total variation gives

$$TV(\rho^*) \leq \liminf_{n\to+\infty} TV(\rho_{\lambda_n}),. \tag{4.19}$$

Accounting for the convexity of K, since it is closed, it is also weakly closed. The function I_K is l.s.c., and

$$I_K(\rho^*) \leq \liminf_{n\to+\infty} I_K(\rho_{\lambda_n}). \tag{4.20}$$

In the same way, Ker$\gamma_{\widetilde{n}_-}$ is weakly closed in $H(v,Q)$, then we deduce that $\rho^* \in H_0 \cap BV(Q)$. Finally, we have, for all $\rho \in H_0 \cap BV(Q)$,

$$\begin{aligned} J(\rho^*) + \lambda TV(\rho^*) + I_K(\rho^*) &\leq \liminf_{n\to+\infty} J(\rho_{\lambda_n}) + (\liminf_{n\to+\infty} \lambda_n)(\liminf_{n\to+\infty} TV(\rho_{\lambda_n})) \\ &\quad + \liminf_{n\to+\infty} I_K(\rho_{\lambda_n}) \\ &\leq \liminf_{n\to+\infty} \left(J(\rho_{\lambda_n}) + \lambda_n TV(\rho_{\lambda_n}) + I_K(\rho_{\lambda_n})\right) \\ &\leq J(\rho) + \lambda TV(\rho) + I_K(\rho), \end{aligned}$$

so $\rho^* = \rho_\lambda$. Moreover, it is clear that ρ_λ is the unique weak limit point of the sequence $(\rho_{\lambda_n})_{n\in\mathbb{N}}$. Assume $T(\lambda_n) = TV(\rho_{\lambda_n}) \underset{n\to+\infty}{\longrightarrow} t$. First it is proved that T has a closed graph.

$$\begin{aligned} J(\rho_\lambda) + \lambda t + I_K(\rho_\lambda) &\leq \liminf_{n\to+\infty} J(\rho_{\lambda_n}) + \lambda t + \liminf_{n\to+\infty} I_K(\rho_{\lambda_n}) \\ &= \liminf_{n\to+\infty} J(\rho_{\lambda_n}) + (\liminf_{n\to+\infty} \lambda_n)(\liminf_{n\to+\infty} TV(\rho_{\lambda_n})) \\ &\quad + \liminf_{n\to+\infty} I_K(\rho_{\lambda_n}) \\ &\leq \liminf_{n\to+\infty} \left(J(\rho_{\lambda_n}) + \lambda_n TV(\rho_{\lambda_n}) + I_K(\rho_{\lambda_n})\right) \\ &\leq J(\rho_\lambda) + \lambda TV(\rho_\lambda) + I_K(\rho_\lambda). \end{aligned}$$

For $\lambda > 0$ fixed, this gives $t \leq TV(\rho_\lambda)$. If $\lambda = 0$, the same inequality holds true. From the decreasing property of T we deduce:

$$TV(\rho_{\lambda_n}) \leq TV(\rho_0), \text{ for all } n > 0. \tag{4.21}$$

Thus, for all $\lambda \geq 0$, we have

$$t \leq TV(\rho_\lambda). \tag{4.22}$$

Gather the l.s.c. of the total variation, with (4.19), we have

$$t = TV(\rho_\lambda), \tag{4.23}$$

which proves that T has a closed graph. Since T is a decreasing function, T is bounded and we conclude that T is a convex closed graph bounded function, thus continuous.
Our concern is now to prove the continuity of Y. Introduce $S : \underset{\lambda}{\mathbb{R}_+} \to \underset{\mapsto J(\rho_\lambda)}{\mathbb{R}_+}$, and arguing as before, it is proved that S is continuous. Recall that ρ_λ is the unique weak limit point of $(\rho_{\lambda_n})_{n\in\mathbb{N}}$, moreover, the continuity of S yields that $\|\rho_{\lambda_n}\|_{H(v,Q)} \underset{n\to+\infty}{\longrightarrow} \|\rho_\lambda\|_{H(v,Q)}$. Consequently, because $H(v,Q)$ is a Hilbert space, the subsequence of $(\rho_{\lambda_n})_{n\in\mathbb{N}}$ that converges weakly to ρ_λ strongly converges to ρ_λ as well. This shows that ρ_λ is the unique limit point of $(\rho_{\lambda_n})_{n\in\mathbb{N}}$. We conclude that $\rho_{\lambda_{n\in\mathbb{N}}}$ strongly converges to ρ_λ. The continuity of Y follows. □

The next result states that by choosing an appropriate λ, it is possible to specify the total variation of the solution to the problem 4.3.

LEMMA 4.3.– *Let ρ_0 be the solution to the problem 4.3 for $\lambda = 0$, and assume that $TV(\rho_0) > 0$. For all $\tau \in (0, TV(\rho_0))$, there exists $\lambda \in \mathbb{R}_+$ and $\rho_\lambda \in H_0 \cap BV(Q)$ solution to the problem 4.3 such that $TV(\rho_\lambda) = \tau$.*

Proof. Let $0 < \tau \in (0, TV(\rho_0))$ be given, Lemma 4.2 claims the existence of $0 < \mu$ sufficiently large such that $T(\mu) - \tau < 0$. Accounting for $0 < T(0) - \tau$, the continuity of function T provides the existence of $0 < \lambda < \mu$ such that $T(\lambda) = TV(\rho_\lambda) = \tau$.

□

We note that the results proved in this section remain true in the discrete setting.

4.1 A Finite element least squares formulation with penalized total variation and nonnegativity

In this section we consider the 2D case. Keeping the same notations as in section 4, the problem (4.3) approximated with a Lagrange finite element method is introduced. For a fixed $\lambda \geq 0$, it reads:

$$u_h = \underset{c_h\in V_h}{\operatorname{argmin}} \quad J(c_h) + \lambda TV(c_h) + I_{K_h}(c_h), \tag{4.24}$$

where $TV(c_h) = \left\| \widetilde{\nabla} c_h \right\|_{L^1(Q)}$, and K_h is defined in (3.7).

A mixed $\mathbb{P}_1/\mathbb{P}_0$ finite element schemes for a primal-dual formulation of the problem has already been proposed in [9]. More precisely, the method is based on the following identity, valid for $u_h \in \mathbb{P}_1$:

$$TV(u_h) = \sup_{q_h\in\mathbb{P}_0,\ |q_h|\leq 1} \int_Q (\nabla u_h \,|\, q_h)\ dx. \tag{4.25}$$

Since this equality is not available for quadrangular meshes anymore (at least for standard $\mathbb{Q}_1/\mathbb{Q}_0$ finite elements), we cannot use this kind of strategy here. So, for numerical convenience, we replace the total

variation by the following discrete operator, that converges to the continuous one when the mesh size tends to zero :

$$TV_d(u) = \sum_{i=1}^{N} \left\| \widetilde{\nabla}_d u(a_i) \right\|_2,$$

where the meshed domain is supposed to be composed of N nodes $a_i = (x_i, y_i)$, $i = 1, \ldots, N$, with

$$\widetilde{\nabla}_d u(a_i) = \left(\nabla_d^1 u(a_i), \nabla_d^2 u(a_i) \right)^t, \quad [4.26]$$

and

$$\|\nabla_d u(a_i)\|_2 = \sqrt{\left(\nabla_d^1 u(a_i)\right)^2 + \left(\nabla_d^2 u(a_i)\right)^2}. \quad [4.27]$$

The discrete operators are defined by:

$$\nabla_d^1 u(a_i) = \begin{cases} \frac{1}{h_x}\left(u(x_i + h_x, y_i) - u(x_i, y_i)\right) & \text{if } x_i < N_1, \\ 0 & \text{if } x_i = N_1, \end{cases} \quad [4.28]$$

and

$$\nabla_d^2 u(a_i) = \begin{cases} \frac{1}{h_y}\left(u(x_i, y_i + h_y) - u(x_i, y_i)\right) & \text{if } x_i < N_2, \\ 0 & \text{if } y_i = N_2, \end{cases} \quad [4.29]$$

with N_1 is the number of rows and N_2 the number of columns of the mesh. So the discrete problem to solve reads: for $0 < \lambda$ given find:

$$u_h = \underset{c_h \in V_h}{\operatorname{argmin}} \quad J(c_h) + \lambda TV_d(c_h) + I_{K_h}(c_h). \quad [4.30]$$

Now consider a matrix form of (4.30). Write $u_h = \sum_{k=1}^{N} U_k \varphi_k$, with $U \in \mathbb{R}^N$. We have: $\widetilde{\nabla}_d U = (\widetilde{\nabla}_d u(a_1), \cdots \widetilde{\nabla}_d u(a_N))^t$, and keeping the same notations as in section 4, define the functions

$$g_1(U) = \frac{1}{2} U^\top A U - U^\top B; \quad g_2(U) = \lambda TV_d(U) + I_{\mathbb{R}_+^N}(U). \quad [4.31]$$

For $0 \leq \lambda$ be fixed, the matrix form discrete problem reads:

$$U = \underset{V \in \mathbb{R}^N}{\operatorname{argmin}} \quad g_1(V) + g_2(V). \quad [4.32]$$

The function g_1 is convex and continuously differentiable, the function g_2 is convex because it is the sum of a continuous convex function (TV_d) and of the indicator function of a closed convex set ($\mathbb{R}_+^N$). Thus we have.

THEOREM 4.4.– *Assume that $\lambda \geq 0$. The problem (4.32) has a unique solution.*

4.2 Algorithms for computing the solution

This subsection is dedicated to algorithms for computing the solution to the problem (4.32). In order to keep the article self contained, it is reminded to the reader that for minimizing a continuously differentiable function h, one step of a gradient method is given by:

$$x_{k+1} = x_k - \sigma \nabla h(x_k);$$
$$x_{k+1} = \arg\min_{x \in \mathbb{R}^N} \frac{1}{2} \|x - x_k + \sigma \nabla h(x_k)\|_2^2 .$$

If we want to minimize the sum of a a continuously differentiable function h plus a convex function Q a combination of a gradient step with a proximal projection can be used:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^N} \frac{1}{2\sigma} \|x - x_k + \sigma \nabla h(x_k)\|_2^2 + Q(x);$$
$$x_{k+1} = prox_{\sigma Q}\left(x_k - \sigma \nabla h(x_k)\right).$$

The FISTA algorithm is a combination of a gradient method with a proximal projection. In our case, since the function g_1 is a continuously differentiable function with Lipschitz continuous gradient, the constant of which is denoted by L, the FISTA algorithm introduced in [10] is well suited for our purposes. It reads:

ALGORITHM 4.1.–

1. *Set* $V_1 = U^0 \in \mathbb{R}^N$, *and* $t_1 = 1$.
2. *For* U^n *and* t^n *given compute up to convergence :*

– $$V^n = \arg\min_{X \in \mathbb{R}^N} \left\{ g_2(X) + \frac{L}{2} \left\| X - \left(U^n - \frac{1}{L} \nabla g_1(U^n) \right) \right\|_2^2 \right\}, \qquad [4.33]$$

– $$t_{n+1} = \frac{1 + \sqrt{1 + 4t_n^2}}{2}, \qquad [4.34]$$

– $$U^{n+1} = V^n + \left(\frac{t_n - 1}{t_{n+1}} \right) \left(V^n - V^{n-1} \right). \qquad [4.35]$$

To compute the solution to the equation (4.33), it expression is split in an other way as previously, making possible an inf-sup formulation of the problem. Set

$$\overline{U} = U^n - \frac{1}{L} \nabla g_1(U^n),$$

the problem (4.33) reads

$$U = \arg\min_{X \in \mathbb{R}^N} \left\{ \frac{L}{2\lambda} \left\| X - \overline{U} \right\|_2^2 + TV_d(X) + I_{K_h}(X) \right\}. \qquad [4.36]$$

Define the functions

$$G(U) = \frac{L}{2\lambda} \left\| U - \overline{U} \right\|_2^2 + I_{K_h}(U),$$

and

$$F(\widetilde{\nabla}_d U) = \left\| \widetilde{\nabla}_d U \right\|_2 = TV_d(U).$$

Thanks to the term $\left\| U - \overline{U} \right\|_2^2$, G is 1-uniformly convex, and, since F is convex and continuous, we know that, for all $P \in \mathbb{R}^N \times \mathbb{R}^N$

$$F(P) = \sup_{P^* \in \mathbb{R}^N \times \mathbb{R}^N} \langle P^*, P \rangle - F^*(P^*), \qquad [4.37]$$

where F^* is the Legendre's transform of F. Moreover, it is well known that $F^* = I_{\mathcal{B}}$, where

$$\mathcal{B} = \left\{ P \in \mathbb{R}^N \times \mathbb{R}^N \mid \|P\|_\infty \leq 1 \right\}.$$

So we have,

$$F(\nabla_d U) = \sup_{P \in \mathbb{R}^N \times \mathbb{R}^N} \langle P, \nabla_d U \rangle - I_{\mathcal{B}}(P).$$

The following primal-dual formulation of problem (4.36) is deduced:

$$\inf_{U \in \mathbb{R}^N} \sup_{P \in \mathbb{R}^N \times \mathbb{R}^N} \langle P, \nabla_d U \rangle + \frac{L}{2\lambda} \left\| U - \overline{U} \right\|_2^2 + I_{K_h}(U) - I_{\mathcal{B}}(P). \qquad [4.38]$$

To solve the saddle point problem (4.38), the following maximizing problem and minimizing problem are alternatively considered:

$$\begin{aligned} \arg\max_{P \in \mathbb{R}^N} &< P, -\nabla_d U > -F^*(P) \\ \arg\min_{U \in \mathbb{R}^N} &< \nabla_d^* P, U > +G(U), \end{aligned}$$

where ∇_d^* is the adjoint operator to the discrete gradient. By using the combination of the gradient decent with a proximal projection reminded at the beginning of this subsection, we have:

$$\begin{aligned} P_{k+1} &= prox_{\sigma F^*}(P_k + \sigma \nabla_d U_k) \\ U_{k+1} &= prox_{\tau G}(U_k - \tau \nabla_d^* P_{k+1}). \end{aligned}$$

In Chambolle and Pock [15], more efficient algorithms based on the same strategy are considered, the convergence of which are proved. More precisely, we use the accelerated algorithm, called *Algorithm 2* in [15].

Then algorithm 4.1 reads:

ALGORITHM 4.2.–

- *Initialization : Set* $V_1 = U^0 \in \mathbb{R}^N$, *and* $t_1 = 1$.

- *For* $n = 0$ *to* $n_{max} - 1$ *: update* $X^n, Y^n, \overline{X}^n$ *as follows :*

  $\mathbf{1}-\ U^{n+\frac{1}{2}} = U^n - \dfrac{1}{L}(AU^n - F)$

  $\mathbf{2}-$ **a**. *For* $\widetilde{L}$ *a Lipschitz constant of* $\widetilde{\nabla}_d$, *choose* $\tau_0, \sigma_0 > 0$, *so that* $\tau_0\sigma_0\widetilde{L}^2 < 1, X^0 \in \mathbb{R}^N, Y^0 \in \mathbb{R}^N \times \mathbb{R}^N$ *and set* $\overline{X}^0 = X^0$.

  b. *For* $k = 0$ *to* $k_{max} - 1$

  $$\begin{array}{l} Y_i^{k+1} = \mathrm{prox}_{\sigma F^*}(Y^n + \sigma\widetilde{\nabla}_d\overline{X}^n), \\ X_i^{k+1} = \mathrm{prox}_{\tau G}(X^n - \tau\widetilde{\nabla}_d^* Y^{n+1}), \\ \theta_k = \dfrac{1}{\sqrt{1+\tau_k}},\ \tau_{k+1} = \theta_k\tau_k,\ \sigma_{k+1} = \dfrac{\sigma_k}{\theta_k} \\ \overline{X}^{k+1} = X^{k+1} + \theta_k(X^{k+1} - X^k) \end{array}$$

  end

  $\mathbf{3}-\ V^n = \overline{X}^{k_{max}}$

  $\mathbf{4}-\ t_{n+1} = \dfrac{1 + \sqrt{1 + 4t_n^2}}{2}$

  $\mathbf{5}-\ U^{n+1} = V^n + \left(\dfrac{t_n - 1}{t_{n+1}}\right)\left(V^n - V^{n-1}\right)$

end

4.3 *Numerical results*

First, the 1D case of a step, moving from left to right without deformations investigated in the previous sections is considered. The computed total variation of the initial signal is 99. In figure 4.3 , the solution, computed with a least squares marching technique after 20 time steps is represented on the left, the $TV_d = 147.51$. On the right, both constraints have been added. For this computed solution $TV_d = 99,01$.

With the next numerical experiment, the influence of parameter λ is studied. The positivity constraint is not used and as expected, the TV constraints acts on oscillations. Unfortunately, if we want to delete oscillations the step is deformed.

To point out the importance of the TV penalization let us come back to the 1D two steps case introduced in figure 3.2. The total variation of the initial signal is 198. The computed solutions are presented after 20 time steps, at the same time in one step case. It can be checked that spurious oscillations disappear with

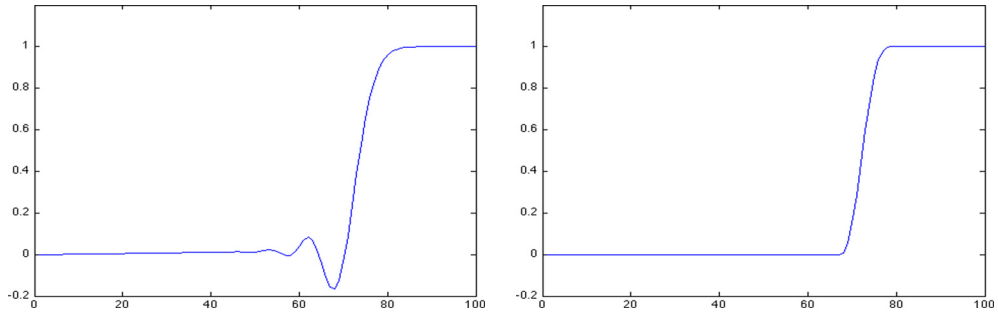

Figure 4.3: *(a) Solution without any constraint; (b) Solution with positivity constraint and TV penalization with $\lambda = 0.001$*

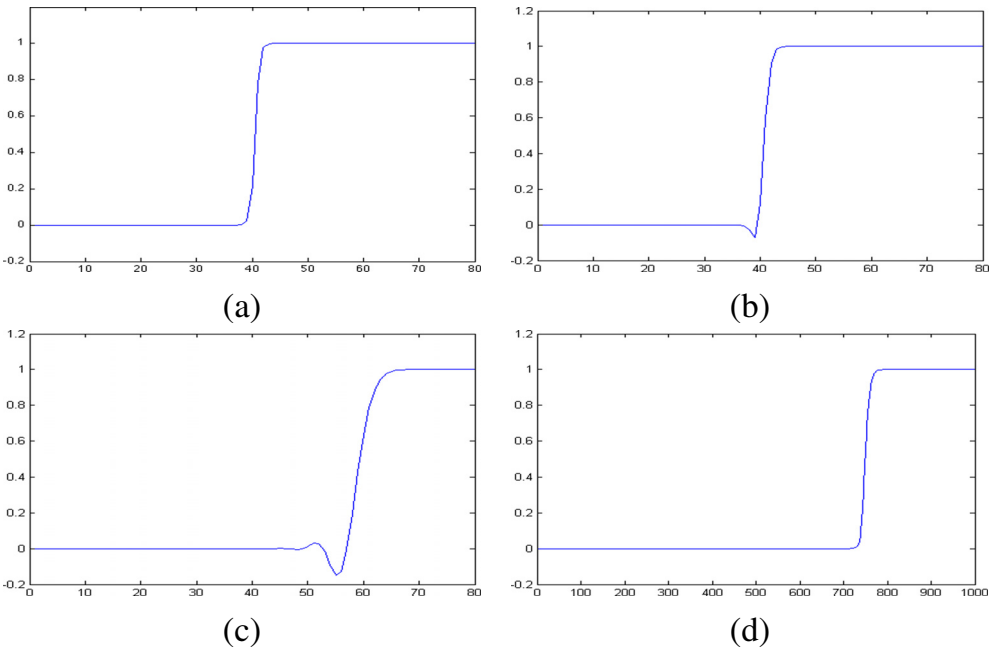

Figure 4.4: *(a) $\lambda = 0.001$; (b) $\lambda = 0.005$; (c) $\lambda = 0.01$; (d) $\lambda = 0.05$.*

positivity and TV constraints, which is here essential. On the left, $TV_d = 281.11$, and $TV_d = 198.03$ on the right.

The next figure shows that combining positivity and TV constraints is the right remedy for working with a very small TV parameter so that the form of the signal is very well preserved. In figure 4.6, we compare two solutions after 20 time steps : on left the solution only with TV penalization, and on right the solution with positivity constraints and TV penalization. We choose for each test a parameter λ so that the total variation of both solutions is 198.035. The parameter is significantly smaller when we use both positivity and T V constraints: $\lambda = 0.004$ than when only a positivity constraint is used: $\lambda = 0.0187$.

The necessity of the TV constraint is clearly shown if we want to remove all the oscillations minimizing modifications of the form of the step.

Before considering 2D examples, let us show that the convergence rate of the finite element method is not affected by the constraints. In the next figure 4.7, on the left, the L^2 error is plotted in logarithmic scale for the least squares marching formulation in red, and in blue, for the formulation where positivity and a total variation constraints have been added. The order of convergence is 2 for both. On the right, the convergence curve of the algorithm 4.1 is given and compared to a straight line, the slope of which is 2.

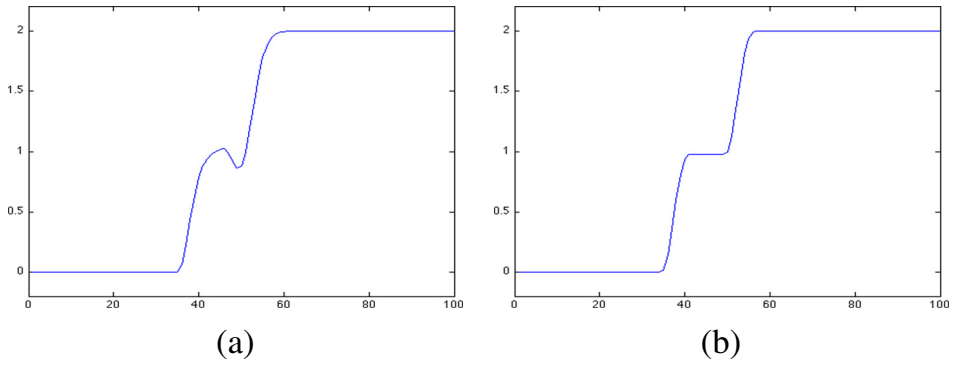

Figure 4.5: *(a) Solution with positivity constraint; (b) Solution with positivity and TV constraints.*

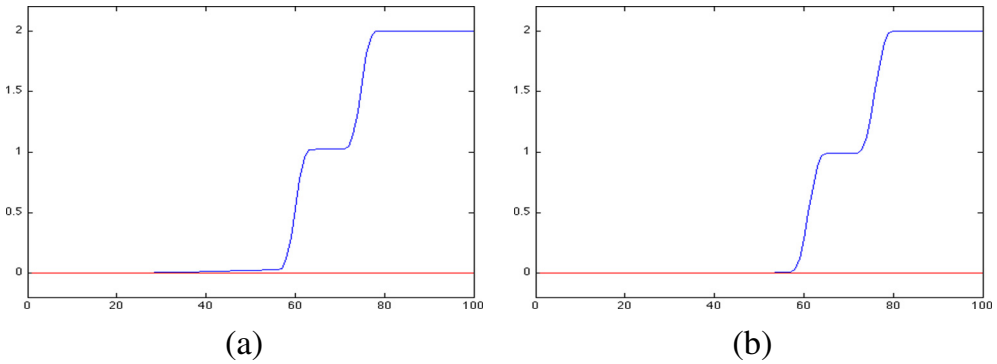

Figure 4.6: *(a) Solution with TV penalization ($\lambda = 0\ 0187$); ($\lambda = 0\ 004$).*
(b) Solution with positivity constraint and TV penalization.

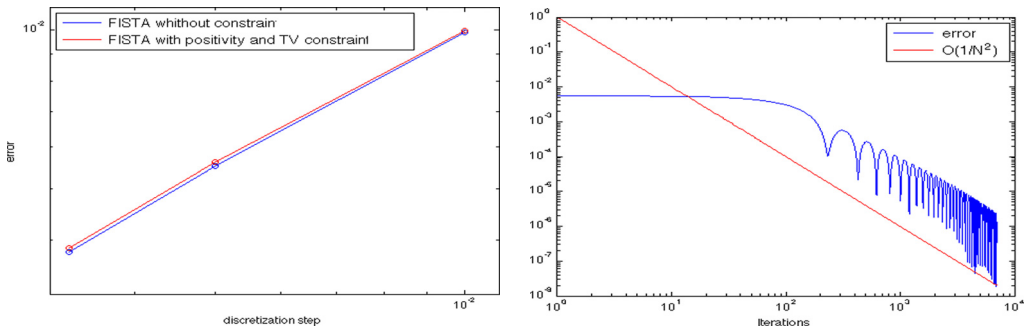

Figure 4.7: *Convergence curves for LS formulation and for the algorithm* 4.1.

A 2D example is now investigated: a bump, the height of which is 250, subjected to a rotating velocity field. The exact solution depicts the bump which rotates without deform when time goes on. In figure 4.8 the velocity field and the initial position of the bump are given.

In figure 4.9 some numerical results are presented for different numerical formulations: least squares (LS in short); LS with extreme values constraint and LS with extreme values and TV constraints. In this example the constraint of the extreme values is realized through the indicator function of the interval $[0, 250]$.The left column depicts a projection on the x, y plan of the bump, then the second column depicts a 3D representation of the bump after a rotation of $\frac{\pi}{2}$, then the two last columns represent slices, respectively, on an horizontal plan going through the center of the bump and on on a vertical plan going through the center of the bump. The first row is concerned with a LS formulation, the second row with a LS formulation with only a positivity constraint and the third row with a LS formulation with positivity and total variation constraints.

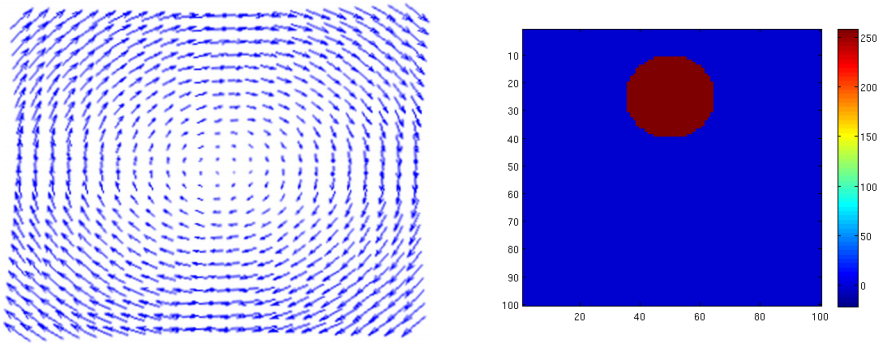

Figure 4.8: *left velocity field, right initial position of the bump*

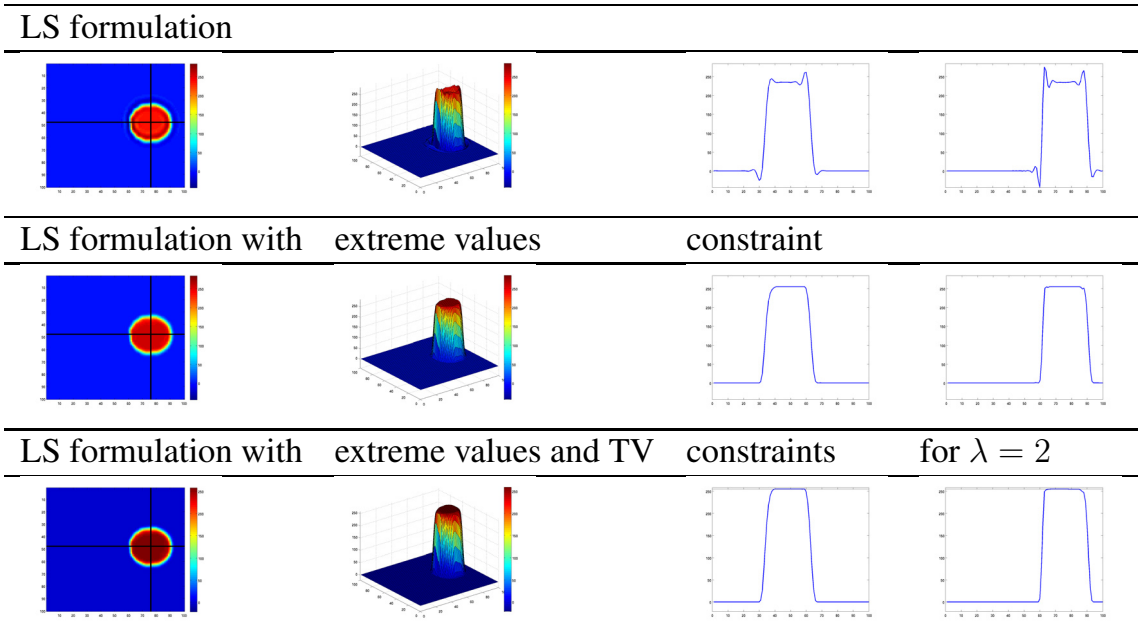

Figure 4.9: *Numerical results with a 2D bump*

If we look carefully at the top of the bump in the left column, one can check that some oscillations (undershoot) subsist for the LS with extreme values constraints, and these oscillations are cancelled with a TV constraint.

Let us come to the 2D double bump case which rotates without deform as in the previous 2D example (figure 4.10). The numerical LS scheme damps a part of the oscillations in the direction of the velocity field, but not in the orthogonal direction. The numerical LS scheme with extreme values and TV constraints handle the oscillations in both directions.

REMARK 4.5.– *As in the 1D case, It can be checked in the previous numerical example that the positivity constraint don't permit to eliminate the intermediate jump oscillations of the signal and justify to use the total variation penalization term in the model.*

In the following table 1 the relations between the values of the computed total variation of the numerical solutions compared to the TV_d value of the initial solution and to the values of λ are reported. One can check the need to be a little more less than the initial TV value if we want to cancel the oscillations.
Let us come to the discussion concerning the extra computational effort due to the constraints. First note that iterative methods are well suited for solving finite element least squares formulation for the transport

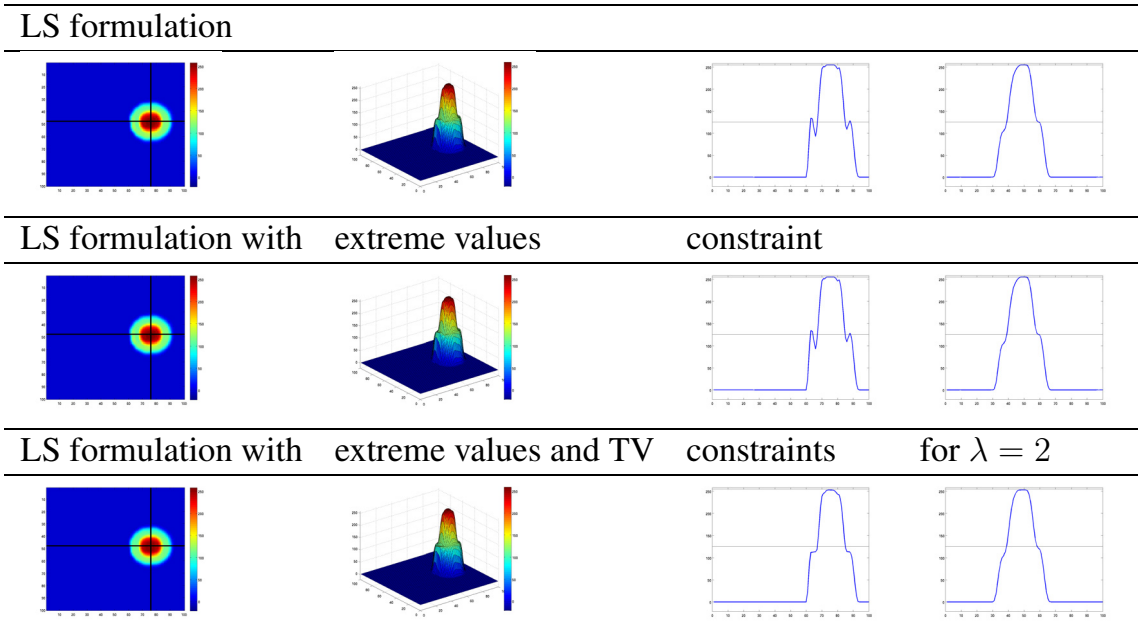

Figure 4.10: *Numerical results with a double bump*

Solution	TV_d	Extreme values constraint	TV_d constraint λ
initial condition	2.0710^4	no	no $\lambda = 0$
LS marching	2.2710^4	no	no $\lambda = 0$
LS marching	1.9310^4	yes	no $\lambda = 0$
LS marching	1.7410^4	yes	yes $\lambda = 2$

Table 1

equation since the condition number of the matrix A is large due to the fact that the diffusion tensor appearing in the variational formulation (see (2.7)) is degenerated [27] and that simple preconditioning factors can be used. In table 2, the computing time for the algorithm 4.1 is compared to the computing time for the same algorithm without any constraint for the 1D two steps signal, with a mesh size of 100 points in space and 50 points in time, that is to say 50 iterations of algorithm 4.1 ($n_{max} = 50$). The positivity constraint does not require computational effort since it consists in evaluating a maximum.

Iterations for computing λ	0	5	10	20	30	40	50
Positivity constraint	No	Yes	Yes	Yes	Yes	Yes	Yes
CPU time (seconds)	1.987	5.568	8.700	15.230	22.343	27.707	34.079

Table 2

In practice, we observe that only a few iterations for the TV loop are needed to have a sufficiently good solution , so that the computing time remains reasonable. In the following example we observe, as illustrated in figure 4.11, that the solution with 10 iterations is very close to the solution with 50 iterations for a CPU time 4 times faster.

Even if in some simple cases, the extreme values control of the computed solution allows to remove the oscillations, a special care has to be taken in the choice of the numerical projection method for ensuring

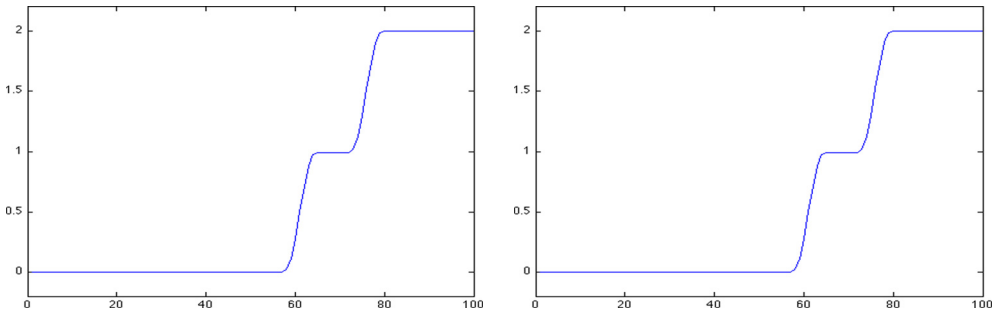

Figure 4.11: *(a) Solution with 10 iterations for the TV loop; (b) solution with 50 iterations for the TV loop.*

this control as it has been proved in section 3. For general situations a TV constraint is necessary. We can conclude that the proposed numerical scheme which controls the extreme values of the solution (through indicator functions) with a penalization of its total variation is able to cancel the spurious oscillations whatever the initial condition is. Moreover, this numerical scheme does not affect the convergence order of the Lagrange finite element method. The proposed method is an acceptable answer to the deficiency of the finite element method for handling the transport equation.

Bibliography

[ALL 92] ALLINEY S., Digital Filters as Absolute Norm Regularizers. *IEEE Transactions on signal processing*, Vol. 40, no. 6, 1992.

[AMB 00] AMBROSIO L., FUSCO N., PALLARA D., *Functions of bounded variation and free discontinuity problems*. Oxford mathematical monographs, Oxford University Press, 2000.

[ATT 06] ATTOUCH H., BUTTAZZO MICHAILLE G., Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization. MPS-SIAM series on optimization, 2006.

[AUB 06] AUBERT G., KORNPROBST P., *Mathematical Problems in Image Processing, Partial Differential Equations and the Calculus of Variations*. Applied Mathematical Sciences 147, Springer Verlag, 2006.

[AUB 99] AUBERT G., DERICHE R., KORNPROBST P., Computing optical flow problem via variational techniques. *SIAM J. Appl. Math.*, 80, 156–182, 1999.

[AZÉ 96] AZÉRAD P., Analyse des équations de Navier-Stokes en bassin peu profond et de l'équation de transport. Thèse de doctorat, Université de Neuchâtel, 1996.

[AZÉ 96] AZÉRAD P., POUSIN J. , Inégalité de Poincaré courbe pour le traitement variationnel de l'équation de transport. C. R. Acad. Sci., Paris, Ser. I **322**, 721–727, 1996.

[AZÉ 96] AZÉRAD P., PERROCHET P., POUSIN J., Space-time integrated least-squares: A simple, stable and precise finite element scheme to solve advection equations as if they were *e*lliptic. M. Chipot, (ed.) et al., Progress in partial differential equations: the Metz surveys 4. Proceedings of the conference given at the University of Metz, France during the 1994–95 'Metz Days'. Harlow: Longman. Pitman Res. Notes Math. Ser. 345, 161–174, 1996.

[BAR 12] BARTELS S., Total Variation Minimization with Finite Elements: Convergence and Iterative Solution, *SIAM J. Numerical Analysis*, **3**, 1162–1180, 2012.

[BEC 09] BECK A., TEBOULLE M., Space-time integrated least squares: a time marching approach. *SIAM J. Img. Sci*., **1**, 183–202, 2009.

[BES 04] BESSON, O., DE MONTMOLLIN, G., A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems, *Int. J. Meth. Fluids*, **44**, 525–543 2004.

[BES 07] BESSON O., POUSIN J., Solutions for Linear Conservation Laws with Velocity Fields in L^∞. *Archive for Rational Mechanics and Analysis*, **186**, 159–175, 2007.

[BOC 09] BOCHEV P.B., GUNZBURGER M.D., *Least-Squares Finite Element Methods*, volume 166, Applied Mathematical Sciences, Springer, 2009.

[BOC 14] BOCHEV P.B., RIDZAL D., PETERSON K., Optimization-based remap and transport: A divide and conqer strategy for feature preserving discretization, *Journal of Computational Physics* **257**, 1113–1139, 2014.

[CHA 11] CHAMBOLLE A., POCK T., A First-Order Primal-Dual Algorithm for Convex Problemswith Applications to Imaging, *Journal of Mathematical Imaging and Vision*, **1**, 120–145, 2011.

[CLA 15] CLARYSSE P., FRIBOULET D. EDITORS, Multi-modality Cardiac Imaging: Processing and Analysis, Digital signal and image processing series, Wiley-ISTE, 2015.

[EVA 09] EVANS J. A., HUGHES T.J.R., SANGALLI G., Enforcement of constraints and maximum principle in the variational multiscale method. *Computer Method in Applied Mechanics and Engineering*, **199** (1–4), 61–76, 2009.

[FRA 91] FRANCA L.P., STENBERG R., Error analysis of Galerkin least squares methods for the elasticity equations. *SIAM-J.-Numer.-Anal*., 28, 1991, pp. 1680–1697.

[FRA 96] FRANCHI B., SERAPIONI R., SERRA CASSANO F., Meyer-Serrin type theorems and relaxation of variational integrals depending on vectors fields. *Houston J. Math.***22**, 859-890, 1996.

[GLO 84] GLOWINSKI R., Numerical Methods for Nonlinear Variational problems. Springer-Verlag, 1984.

[GUE 04] GUERMONT J. L., A finite element technique for solving first-order PDEs in L^p. *SIAM journal of numerical analysis*. **42** (2), 714–737, 2004.

[GUE 08] GUERMONT J. L., MARPEAU F. , POPOV B., A fast algorithm for solving first-order PDEs by L^1 minimization. *COMMUN. MATH. SCI*. Vol. 6 No 1., pp. 199–216, 2008.

[JIA 93] JIANG B.-N., Non-oscillatory and Non-diffusive Solution of Convection Problem by the Iterativeley Reweighted Least-Squares Finite Element Method. *Journal of computational physics*, **105**, 108–121, 1993.

[KUZ 12] KUZMIN D., Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes. *JCAM* **236**, 2317–2337, 2012.

[LIO 96] LIONS P.L., *Mathematical topics in fluid mechanics*. Vol. 1; Oxford Science Publication, Calderon Press, 1996.

[LIS 08] LISKA R., SHASHKOV M., Enforcing the discrete maximum principle for linear finite element solutions of second-order elliptic problems. *Com. Comput. Phys*., **3** (4), 852–877, 2008.

[MON 01] DE MONTMOLLIN G., Méthode STILS pour l'équation de transport: comparaisons et analyses. Etude d'un modèle de fermeture pour la loi de Darcy. Thèse de doctorat, Université de Neuchâtel, 2001.

[ROK 08] ROKAFELLAR R.T., WETS R.J.B., Variational analysis. Springer, 2008.

[STE 04] DE STERCK H., THOMAS A., MANTEUFFEL A., STEPHEN F., MACCORMIK L. OLSON, Least-squares finite element methods and algebraic multigrid solvers for linear hyperbolic PDEs. *SIAM J. Sci Comput*. Vol. 26, **No. 1**, pp. 31–54, 2004.

[YOU 08] YOUNES L., ARRATE F., MILLER M. , Evolution Equations in Computational Anatomy. *Neuroimage*, 2008, pp. S40–S50.

