A numerical procedure and unified formulation for the adjoint approach in hyperbolic PDE-constrained optimal control problems

GINO I. MONTECINOS*

Centro de Modelamiento Matemático (CMM), Universidad de Chile,

Beauchef 851, Torre Norte, Piso 7, Santiago, Chile

JUAN C. LÓPEZ-RÍOS

Escuela de Ciencias Matemáticas y Tecnología Informática, YACHAY TECH,

San Miguel de Urcuquí, Hacienda San José s/n, Ecuador

JAIME H. ORTEGA

Centro de Modelamiento Matemático (CMM) and Departamento de Ingeniería Matemática, Universidad de Chile,

Beauchef 851, Torre Norte, Piso 5, Santiago, Chile

RODRIGO LECAROS

Departamento de Matemática, Universidad Técnica Federico Santa María,

Casilla 110-V, Valparaíso, Chile

The present paper aims at providing a numerical strategy to deal with PDE-constrained optimization problems solved with the adjoint method. It is done through out a unified formulation of the constraint PDE and the adjoint model. The resulting model is a non-conservative hyperbolic system and thus a finite volume scheme is proposed to solve it. In this form, the scheme sets in a single frame both constraint PDE and adjoint model. The forward and backward evolutions are controlled by a single parameter η and a stable time step is obtained only once at each optimization iteration. The methodology requires the complete eigenstructure of the system as well as the gradient of the cost functional. Numerical tests evidence the applicability of the present technique.

Keywords: The adjoint method, PDE-constrained optimal control, hyperbolic conservation laws.

1. Introduction

We are concerned with PDE-constrained optimal control problems. In general optimization problems are made up of the following ingredients; i) state variables, ii) design parameters, iii) objectives or cost functionals and iv) constraints that candidate state and design parameters are required to satisfy. The optimization problem is then to find state and design parameters that minimize the objective functional subject to the requirement that constraints are satisfied, see [3] for further details. PDE-constrained optimization problems are those where the constraint consists of partial differential equations. These optimization problems arise naturally in control theory and inverse problems. Are usually employed as a methodology to obtain an approximate solution often from numerical simulations, see [21, 20] to mention but a few. Mathematical properties of these problems, as existence and ill-possedness, have been widely studied in the literature, see for instance [1, 15] and they will not be of interest in this

work. In this paper, we are interested on generating a procedure to obtain numerical solutions to be as general as possible. There are several successful strategies to solve these types of minimization problems: i) the one-shot or Lagrange multiplier method ii) based on sensitivity equations and iii) based on adjoint equations. Of interest in this work is the strategy based on adjoint equations. The interested reader can find further information regarding approaches i) and ii) in [3, 13] and references therein. One-shot methods cannot be used to solve most practical optimization problems especially involving time-dependent partial differential equations. Methods ii) and iii) are iterative processes involving the gradient of the objective functional, since the functional involves the solution of partial differential equations the task of obtaining the gradient becomes cumbersome, thus a safe option is just the use of approximations of them. However, in these problems even the approximation of gradients is a very challenging issue. Sensitivity equations require the differentiation of the state equation with respect to the design parameters. The simple form to achieve this is using a difference quotient approximation similarly to finite difference approximation as suggested in [3]. This strategy can be successfully applied for steady state models and for a finite and small number of design parameters. It becomes prohibitive for time-dependent cases. In this sense, the adjoint method is the suitable alternative.

The adjoint method consists of the use of a linearized equation derived from the constraint PDE, it provides a new set of adjoint states, and the gradient of the cost functional can be completely determined in terms of the state variables, design parameters and adjoint states. The equation for the state variables, as well as for the adjoint parameter, is solved only once. It makes the adjoint method to be more efficient than the sensitivity method, see [15] for a detailed discussion about the performance of both approaches. In this work, we are going to consider the adjoint approach for solving optimization problems related to hyperbolic type PDE's. Additionally, for hyperbolic equations, the design parameters, which are parameters of models, can be incorporated into the governing equation as a new variable, and then the problem of finding parameters can be cast into a problem of finding the initial condition.

The constraint PDE normally is written in a conservation form and thus suitable schemes are those corresponding to the family of conservative methods. On the other hand, the adjoint models are linearization of the constraint PDE, but this is a quasilinear type model for adjoint variables and thus they are also of hyperbolic type, but these models are non-conservative in a mathematical sense. It is a very important issue from a numerical point of view, the right description of the weave propagation has to be dealt in the frame of the path conservative methods for partial differential equations involving non-conservative products, see [32, 7] for a detailed study of non-conservative partial differential equations with a particular emphasis on hyperbolic types.

As the constraint PDE and the adjoint model are both hyperbolic, they can be written as a unified model incorporating both the constraint PDE and the adjoint model in a single system of balance laws. In this paper, we back up on the unified formulation to set a numerical methodology able to simultaneously deal with the conservative form of constraint PDE and the adjoint model. The aim is to provide a general frame to deal with optimization problems through the adjoint method. Moreover, the present approach works even if the constraint PDE is non-conservative. In the literature, there exist several works dealing with minimization considering hyperbolic conservation laws, see [18, 17, 5, 25, 39] to mention but a few. However, to best of our knowledge, the present approach has not been reported previously and this incorporates in a natural way the non-conservative structure of the adjoint model.

On the other hand, regarding the solution of hyperbolic conservation laws, one of the successful methods are those based on the finite volume approach, which makes use of numerical fluxes at the interfaces of computational cells, providing a compact one-step evolutionary formula. The accuracy of the approximation via this approach depends on the degree of resolution of numerical fluxes. One approach for achieving high resolution is by mean of the use of the so called Generalized Riemann

Problem (GRP), see [6, 27] for further details and comparison among existing GRP solvers. These are building blocks of a class of high-order numerical methods known as ADER schemes [37, 38]. The numerical scheme presented in this work belong to the class of ADER methods. The scheme uses a modified version of the Osher-Solomon Riemann solver [31] presented by Dumber and Toro, [9], which is able to account for the non-conservative terms, using the so-called path conservative approach, it is a second order method for hyperbolic problems in conventional finite volume setting and it inherits the properties of stability, well-balance, consistency and robustness. However, the implementation has to be adapted for accounting the correct wave propagation for the evolution of both, the PDE constraint and the adjoint model.

In order to assess the performance of the present method, we carry out comparisons between the present scheme and a conventional method based on [26] for solving PDE-constrained optimization problems. We adapt the scheme in [26] implementing an upwind type scheme for the solution of the constraint and for the adjoint model, the same finite difference approach is kept but it is carried out in an explicit time evolution to be consistent with the present method which is globally explicit in time.

The rest of the paper is organized as follows. In Section 2. the unified formulation of the primal dual strategy for a general system of conservation laws is presented, after the formal derivation of the adjoint system. We also state a general theorem providing the existence of classical solutions for the unified system and study a related optimal control problem. In Section 3., a numerical scheme for solving conservative and non-conservative hyperbolic laws is presented and, as a consequence, a scheme for solving the unified primal-adjoint system. In Section 4. some specific numerical examples are presented using the numerical scheme from Section 3.. Finally in Section 5. some conclusions are drawn.

2. PDE-constrained optimization problem

The purpose of this section is to provide a unified formulation of the primal dual strategy, raised from the study of certain optimal control problems concerning the finding of a parameter on a general conservation law.

First, we define the hyperbolic PDE to be considered and we prove that the finding of a parameter on this system is reduced to find an initial condition. Next we compute the adjoint system and set the unified formulation. Then we provide some conditions to establish the existence and uniqueness of classical solutions of the unified system. Finally we set the control problem and prove the existence of an optimal solution.

We consider the system

$$\left.\begin{array}{c} \partial_t \mathbf{U} + \partial_x \mathbf{R}(\mathbf{U}, \mathbf{b}) = \mathbf{L}(\mathbf{U}, \mathbf{b}) + \tilde{\mathbf{B}}(\mathbf{U}, \mathbf{b})\partial_x \mathbf{b} , \ t \in [0, T] , \\ \mathbf{U}(x, 0) = \mathbf{H}(x) , \end{array}\right\} \qquad (2..1)$$

where $T > 0$ is a finite time, $\mathbf{H}(x) \in \mathbb{R}^m$ is a prescribed initial condition, $\mathbf{U} \in \mathbb{R}^m$ is a vector of states, $\mathbf{R}(\mathbf{U}, \mathbf{b}) \in \mathbb{R}^m$ a flux function, $\mathbf{L}(\mathbf{U}, \mathbf{b}) \in \mathbb{R}^m$ is a source function, $\mathbf{b}(x) \in \mathbb{R}^n$ is a vector accounting for model parameters and the expression $\tilde{\mathbf{B}}(\mathbf{U}, \mathbf{b})\partial_x \mathbf{b}$ represents source terms which may also include the influence of parameter derivatives. Here, parameters $\mathbf{b}$ only depend on space. Since one main goal is to provide a strategy for finding $\mathbf{b}$ as the minimum of a given functional J, let us write J as

$$J(\mathbf{U}, \mathbf{b}) = \frac{1}{2} \int_0^T \int_{\mathbb{R}} |\psi(\mathbf{U}, \mathbf{b}) - \bar{\psi}|^2 dx dt + \frac{\alpha}{2} \|\mathbf{b}\|_{H^k}^2, \qquad (2..2)$$

where $\psi(\mathbf{U}, \mathbf{b})$ is some scalar field, representing some measurement operator, similarly $\bar{\psi}$ is some available measurement, α is a non-negative constant and k is an exponent to be determined.

In an optimization setting we identify the state variables $\mathbf{U}$, design parameters $\mathbf{b}$, objectives or cost functional given by (2..2) and constraints consisting of (2..1) that candidate state $\mathbf{U}$ and design parameters $\mathbf{b}$ are required to satisfy.

Notice that, parameters $\mathbf{b}$ only depend on space, then $\partial_t \mathbf{b}(x) = 0$. So, we can include $\mathbf{b}$ as a new variable of system (2..1), as follows

$$\left. \begin{array}{c} \partial_t \mathbf{U} + \partial_x \mathbf{R}(\mathbf{U},\mathbf{b}) = \mathbf{L}(\mathbf{U},\mathbf{b}) + \tilde{\mathbf{B}}(\mathbf{U},\mathbf{b})\partial_x \mathbf{b} \,, \\ \\ \partial_t \mathbf{b} = \mathbf{0} \,. \end{array} \right\} \tag{2..3}$$

So, written in the vector form, (2..3) becomes

$$\partial_t \tilde{\mathbf{U}} + \partial_x \tilde{\mathbf{R}}(\tilde{\mathbf{U}}) + \tilde{\mathbf{M}}\partial_x \tilde{\mathbf{U}} = \tilde{\mathbf{L}}(\tilde{\mathbf{U}}) \,, \tag{2..4}$$

where

$$\begin{array}{c} \tilde{\mathbf{U}} = \left[\begin{array}{c} \mathbf{U} \\ \mathbf{b} \end{array} \right] \,, \ \tilde{\mathbf{R}}(\tilde{\mathbf{U}}) = \left[\begin{array}{c} \mathbf{R}(\mathbf{U},\mathbf{b}) \\ \mathbf{0} \end{array} \right] \,, \\ \\ \tilde{\mathbf{L}}(\tilde{\mathbf{U}}) = \left[\begin{array}{c} \mathbf{L}(\mathbf{U},\mathbf{b}) \\ \mathbf{0} \end{array} \right] \,, \ \tilde{\mathbf{M}}(\tilde{\mathbf{U}}) = \left[\begin{array}{cc} \mathbf{0} & -\tilde{\mathbf{B}}(\mathbf{U},\mathbf{b}) \\ \mathbf{0} & \mathbf{0} \end{array} \right] \,. \end{array} \tag{2..5}$$

In this new system, $\mathbf{b}(x)$ is completely fixed once an initial condition is provided, so the task of finding parameters in a model system as (2..1) is equivalent to find the initial condition of the enlarged model system (2..3), henceforth referred to as *constraint PDE*. From now on, instead of system (2..1) we are going to consider system

$$\left. \begin{array}{c} \partial_t \tilde{\mathbf{U}} + \partial_x \tilde{\mathbf{R}}(\tilde{\mathbf{U}}) + \tilde{\mathbf{M}}\partial_x \tilde{\mathbf{U}} = \tilde{\mathbf{L}}(\tilde{\mathbf{U}}) \,, \ t \in [0,T] \,, \\ \tilde{\mathbf{U}}(x,0) = \tilde{\mathbf{H}}(x) \,, \end{array} \right\} \tag{2..6}$$

where $\tilde{\mathbf{H}}(x) = [\mathbf{H},\mathbf{0}]^T$. The following result provides the conditions on system (2..1) which turns system (2..6) to be hyperbolic.

PROPOSITION 2..1 Let (2..1) be a hyperbolic system in the variable $\mathbf{U}$, with $\{\lambda_1,...,\lambda_m\}$ the corresponding eigenvalues of $\partial \mathbf{R}/\partial \mathbf{U}$, where $\lambda_i \neq 0$. Then (2..6) is also a hyperbolic system in the variables $\tilde{\mathbf{R}}(\tilde{\mathbf{U}}) = [\mathbf{U},\mathbf{b}]^T$. The set of eigenvalues is given by

$$\{\lambda_1,...,\lambda_m,\underbrace{0,...,0}_{n \ times}\}$$

and the corresponding eigenvectors are given by

$$\{[\mathbf{v}_1,\mathbf{0}_n]^T,...,[\mathbf{v}_m,\mathbf{0}_n]^T,[\tilde{\mathbf{V}}_1,\mathbf{e}_1]^T,...,[\tilde{\mathbf{V}}_n,\mathbf{e}_n]^T\} \,,$$

with $\tilde{\mathbf{V}}_j = -(\partial \mathbf{R}/\partial \mathbf{U})^{-1}(\partial \mathbf{R}/\partial \mathbf{b} - \tilde{\mathbf{B}})_j$, where $(\partial \mathbf{R}/\partial \mathbf{b} - \tilde{\mathbf{B}})_j$ is the $j-th$ column vector of the matrix $(\partial \mathbf{R}/\partial \mathbf{b} - \tilde{\mathbf{B}})$. Here, $\mathbf{v}_i$ is an eigenvector associated with λ_i, $\mathbf{0}_m$ is the zero vector in $\mathbb{R}^m$, $\mathbf{e}_i$ is the ith canonical vector of $\mathbb{R}^n$.

Proof. The Jacobian matrix of system (2..6) is given by

$$\mathbf{J_R} := \partial \tilde{\mathbf{R}}/\partial \tilde{\mathbf{U}} + \tilde{\mathbf{M}} = \left[\begin{array}{cc} \partial \mathbf{R}/\partial \mathbf{U} & (\partial \mathbf{R}/\partial \mathbf{b} - \tilde{\mathbf{B}}) \\ \mathbf{0} & \mathbf{0} \end{array} \right] \,. \tag{2..7}$$

Therefore

$$det(\mathbf{J_R} - \lambda \mathbf{I}_{n+m}) = -\lambda^n det(\partial \mathbf{R}/\partial \mathbf{U} - \mathbf{I}_m) = -\lambda^n p(\lambda) \,, \tag{2..8}$$

where $\mathbf{I}_{n+m}$ and $\mathbf{I}_m$ are the identity matrix in $\mathbb{R}^{n+m}$ and $\mathbb{R}^m$, respectively. Here, $p(\lambda)$ is the characteristic polynomial of $\partial \mathbf{R}/\partial \mathbf{U} - \mathbf{I}_m$ which has m roots, because (2..1) is hyperbolic. This proves

$$\{\lambda_1, ..., \lambda_m, \underbrace{0, ..., 0}_{n\ times}\}$$

are the eigenvalues of $\mathbf{J_R}$.

To complete the proof, we must obtain a set of $m+n$ linearly independent eigenvectors. That means, for each $\mu \in \{\lambda_1, ..., \lambda_m, \underbrace{0, ..., 0}_{n\ times}\}$ we look for $\mathbf{W}$ such that

$$\mathbf{J_R}\mathbf{W} = \mu \mathbf{W} \,. \tag{2..9}$$

Here, we obtain eigenvectors of the form $\mathbf{W} = [\tilde{\mathbf{V}}, \tilde{\mathbf{W}}]^T$, so in a block-wise fashion we look for vectors $\tilde{\mathbf{V}}$ and $\tilde{\mathbf{W}}$ such that

$$\left. \begin{array}{r} (\partial \mathbf{R}/\partial \mathbf{U})\tilde{\mathbf{V}} + (\partial \mathbf{R}/\partial \mathbf{b} - \tilde{\mathbf{B}})\tilde{\mathbf{W}} = \mu \tilde{\mathbf{V}} \,, \\ \mathbf{0} = \mu \tilde{\mathbf{W}} \,. \end{array} \right\} \tag{2..10}$$

Notice that, setting $\mu = \lambda_i$ and $\tilde{\mathbf{V}} = \mathbf{v}_i$, then (2..10) is satisfied for $\tilde{\mathbf{W}} = \mathbf{0}_n$. It provides m independent vectors $\{[\mathbf{v}_1, \mathbf{0}_n]^T, ..., [\mathbf{v}_m, \mathbf{0}_n]^T\}$. On the other hand, if $\mu = 0$ we obtain

$$(\partial \mathbf{R}/\partial \mathbf{U})\tilde{\mathbf{V}} + (\partial \mathbf{R}/\partial \mathbf{b} - \tilde{\mathbf{B}})\tilde{\mathbf{W}} = \mathbf{0} \,. \tag{2..11}$$

If we set $\tilde{\mathbf{W}} = \mathbf{e}_j$, then

$$(\partial \mathbf{R}/\partial \mathbf{U})\tilde{\mathbf{V}} = -(\partial \mathbf{R}/\partial \mathbf{b} - \tilde{\mathbf{B}})_j \,, \tag{2..12}$$

where $(\partial \mathbf{R}/\partial \mathbf{b} - \tilde{\mathbf{B}})_j$ is the $j-th$ column vector of the matrix $(\partial \mathbf{R}/\partial \mathbf{b} - \tilde{\mathbf{B}})$. Since $\partial \mathbf{R}/\partial \mathbf{U}$ is invertible, there exists $\tilde{\mathbf{V}}_j \in \mathbb{R}^m$ such that

$$\tilde{\mathbf{V}}_j = -(\partial \mathbf{R}/\partial \mathbf{U})^{-1}(\partial \mathbf{R}/\partial \mathbf{b} - \tilde{\mathbf{B}})_j \,.$$

Then the proof holds. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \square$

REMARK 2..1 Notice that the previous result ensure that extended system (2..6) is hyperbolic, whenever system (2..1) is. However, it requires (2..1) to have an invertible Jacobian matrix. It is not guaranteed in the general case. However, the cases of interest in this study do so. Otherwise, the verification has to be done case by case.

As a consequence, the problem of finding a parameter vector for system (2..1) becomes an inverse problem where the aim is to find an initial condition for $\mathbf{b}$ which is a conventional variable in a system of hyperbolic balance laws.

2.1. *The adjoint method for solving PDE-constrained optimization problems*

In this section we derive a system of first order optimality conditions for system (2..6), subject to the functional (2..2) with $\alpha = 0$. Even though these calculations are formal, we will provide the corresponding analysis in sections 2.3. and 2.4.. At this point our main interest is to present the unified primal-dual formulation and the main issues related with. Since it is well known for this strategy [15], the so-called adjoint state is, in some sense, a linearization of the constraint PDE. The dependence of the sought parameter **b** in (2..6) is implicit but, as was mentioned above, can be seen as the tracking of an initial condition.

As we did in Proposition (2..1), for $i = 1 : m + n$, let us write (2..6) as

$$\left. \begin{array}{c} \partial_t \tilde{\mathbf{U}}_i + \sum_{j=1}^{m+n} \mathbf{J}_{\mathbf{R}i,j} \partial_x \tilde{\mathbf{U}}_j = \tilde{\mathbf{L}}_i \ , \ t \in [0,T] \ , \\ \tilde{\mathbf{U}}_i(x,0) = \tilde{\mathbf{H}}_i(x) \ . \end{array} \right\} \tag{2..13}$$

Then

$$\partial_t \delta \tilde{\mathbf{U}}_i + \sum_{j=1}^{m+n} \mathbf{J}_{\mathbf{R}i,j} \partial_x \delta \tilde{\mathbf{U}}_j + \sum_{j=1}^{m+n} \sum_{k=1}^{m+n} \frac{\partial \mathbf{J}_{\mathbf{R}i,j}}{\partial_x \tilde{\mathbf{U}}_k} \delta \tilde{\mathbf{U}}_k \partial_x \tilde{\mathbf{U}}_j = \\ \sum_{k=1}^{m+n} \frac{\partial \tilde{\mathbf{L}}_k}{\partial \tilde{\mathbf{U}}_\mathbf{k}} \delta \tilde{\mathbf{U}}_k \ . \tag{2..14}$$

If we denotes the inner product in $L^2(\mathbb{R})$ by $\langle \cdot, \cdot \rangle$ and multiply last equation by functions P_i, such that $lim_{|x| \to \infty} P_i(x,t) = 0$, $\forall \, t$, after integrating in space and time, we obtain

$$\int_0^T \langle \partial_t \delta \tilde{\mathbf{U}}_i, P_i \rangle + \sum_{j=1}^{m+n} \int_0^T \langle \partial_x \delta \tilde{\mathbf{U}}_j, \mathbf{J}_{\mathbf{R}i,j} P_i \rangle \\ + \sum_{j=1}^{m+n} \sum_{k=1}^{m+n} \int_0^T \langle \delta \tilde{\mathbf{U}}_k, \frac{\partial \mathbf{J}_{\mathbf{R}i,j}}{\partial \tilde{\mathbf{U}}_k} \partial_x \tilde{\mathbf{U}}_j P_i \rangle = \\ \sum_{k=1}^{m+n} \int_0^T \langle \delta \tilde{\mathbf{U}}_k, \frac{\partial \tilde{\mathbf{L}}_k}{\partial \tilde{\mathbf{U}}_\mathbf{k}} P_i \rangle \ . \tag{2..15}$$

Summing on indices i, yields

$$\int_0^T \sum_{i=1}^{m+n} \langle \partial_t \delta \tilde{\mathbf{U}}_i, P_i \rangle - \sum_{j=1}^{m+n} \int_0^T \langle \delta \tilde{\mathbf{U}}_j, \sum_{i=1}^{m+n} \partial_x (\mathbf{J}_{\mathbf{R}i,j} P_i) \rangle \\ + \sum_{j=1}^{m+n} \sum_{k=1}^{m+n} \int_0^T \langle \delta \tilde{\mathbf{U}}_k, \sum_{i=1}^{m+n} \frac{\partial \mathbf{J}_{\mathbf{R}i,j}}{\partial \tilde{\mathbf{U}}_k} \partial_x \tilde{\mathbf{U}}_j P_i \rangle = \\ \sum_{k=1}^{m+n} \int_0^T \langle \delta \tilde{\mathbf{U}}_k, \sum_{i=1}^{m+n} \frac{\partial \tilde{\mathbf{L}}_k}{\partial \tilde{\mathbf{U}}_\mathbf{k}} P_i \rangle \ . \tag{2..16}$$

Then integrating by parts, we obtain

$$\sum_{k=1}^{m+n} \langle \delta \tilde{\mathbf{U}}_k, P_k \rangle |_0^T - \int_0^T \sum_{k=1}^{m+n} \langle \delta \tilde{\mathbf{U}}_k, \partial_t P_k \rangle - \sum_{k=1}^{m+n} \int_0^T \langle \delta \tilde{\mathbf{U}}_k, \sum_{i=1}^{m} \mathbf{J}_{\mathbf{R}i,k} \partial_x P_i \rangle \\ - \sum_{k=1}^{m+n} \int_0^T \langle \delta \tilde{\mathbf{U}}_k, \sum_{i=1}^{m} \mathbf{D}_{k,j} \partial_x \tilde{\mathbf{U}}_j \rangle = \sum_{k=1}^{m+n} \int_0^T \langle \delta \tilde{\mathbf{U}}_k, \sum_{i=1}^{m+n} \frac{\partial \tilde{\mathbf{L}}_k}{\partial \tilde{\mathbf{U}}_\mathbf{k}} P_i \rangle \ , \tag{2..17}$$

where

$$\mathbf{D}_{k,j} = \sum_{i=1}^{m+n} \left(\frac{\partial \mathbf{J}_{\mathbf{R}i,k}}{\partial \tilde{\mathbf{U}}_j} - \frac{\partial \mathbf{J}_{\mathbf{R}i,j}}{\partial \tilde{\mathbf{U}}_k} \right) P_i \ . \tag{2..18}$$

On the other hand, $\delta J(\mathbf{b}) = \int_0^T \int_{-\infty}^{\infty} (\psi(\mathbf{U}, \mathbf{b}) - \bar{\psi})(\nabla_{\mathbf{U}} \psi^T \delta \mathbf{U} + \nabla_{\mathbf{b}} \psi^T \delta \mathbf{b})$ or

$$\delta J(\mathbf{b}) = \sum_{k=1}^{m+n} \int_0^T \langle \delta \tilde{\mathbf{U}}_k, (\psi - \bar{\psi}) \frac{\partial \psi}{\partial \tilde{\mathbf{U}}_k} \rangle .$$

Moreover

$$\delta J = \sum_{j=1}^{n} \langle \nabla J, \delta \mathbf{b} \rangle , \tag{2..19}$$

so, by identifying terms we obtain

$$\nabla J_j(x) = P_{m+j}(x, 0) , \tag{2..20}$$

for all $j = 1, ..., n$. Imposing $P_i(x, T) = 0$ for $i = 1, ..., n+m$ and $P_i(x, 0) = 0$ for $i = 1, ..., m$ we obtain the adjoint problem

$$\partial_t P_k + \sum_{i=1}^{m+n} \mathbf{J_R}_{i,k} \partial_x P_i + \sum_{i=1}^{m+n} D_{k,j} \partial_x \tilde{\mathbf{U}}_j = \sum_{i=1}^{m+n} \frac{\partial \tilde{L}_i}{\partial \tilde{\mathbf{U}}_k} P_i + (\psi - \bar{\psi}) \frac{\partial \psi}{\partial \tilde{\mathbf{U}}_k} . \tag{2..21}$$

In a matrix form

$$\partial_t \mathbf{P} + \mathbf{J_R}^T \partial_x \mathbf{P} = \tilde{S}(\tilde{\mathbf{U}}, \mathbf{P}) ,$$
$$\mathbf{P}(x, T) = \mathbf{0} , \tag{2..22}$$

where $\frac{1}{2} \nabla_{\tilde{\mathbf{U}}}((\psi - \bar{\psi})^2)_k = (\psi - \bar{\psi}) \frac{\partial \psi}{\partial \tilde{\mathbf{U}}_k}$ and $\mathbf{P} = [P_1, ..., P_{n+m}]^T$. The source term in this case has the form

$$\tilde{S}(\tilde{\mathbf{U}}, \mathbf{P}) = \frac{\partial \tilde{L}}{\partial \tilde{\mathbf{U}}}^T \mathbf{P} + \frac{1}{2} \nabla_{\tilde{\mathbf{U}}}((\psi - \bar{\psi})^2) - \mathbf{D} \partial_x \tilde{\mathbf{U}} .$$

In the sequel, we shall refer to $\mathbf{P}$ as *the adjoint state*. Notice that system (2..22) is also hyperbolic, with the same eigenvalues that system (2..13). It is because the Jacobian matrix of the system (2..22) is $\mathbf{J}_R^T$. Moreover, notice that system (2..22), henceforth referred to as the adjoint model, evolves back in time from T to 0.

2.2. *The unified formulation*

In this section, we construct a model system which accounts for both the PDE constraint and the adjoint model, aimed to construct a numerical method, able to solve simultaneously both systems including the treatment of non-conservative products. So, let us consider the following unified primal-dual system

$$\partial_t \mathbf{Q} + \partial_x \mathbf{F}(\mathbf{Q}) + \mathbf{B}(\mathbf{Q}) \partial_x \mathbf{Q} = \mathbf{S}(\mathbf{Q}) , \tag{2..23}$$

where

$$\mathbf{Q} = \begin{pmatrix} \tilde{\mathbf{U}} \\ \mathbf{P} \end{pmatrix} , \quad \mathbf{F}(\mathbf{Q}) = \begin{pmatrix} \tilde{\mathbf{R}}(\tilde{\mathbf{U}}) \\ \mathbf{0} \end{pmatrix} ,$$
$$\mathbf{B}(\mathbf{Q}) = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{D} & \mathbf{J_R}^T \end{pmatrix} , \quad \mathbf{S}(\mathbf{Q}) = \begin{pmatrix} \tilde{\mathbf{L}} \\ -\frac{\partial \tilde{\mathbf{L}}}{\partial \tilde{\mathbf{U}}}^T \mathbf{P} + \frac{1}{2} \nabla_{\tilde{\mathbf{U}}}((\psi - \bar{\psi})^2) \end{pmatrix} . \tag{2..24}$$

Notice that the state $\mathbf{Q} \in \mathbb{R}^{2(m+n)}$ contains the state variable $\mathbf{U} \in \mathbb{R}^m$, the model parameter $\mathbf{b} \in \mathbb{R}^n$ and the adjoint state $\mathbf{P} \in \mathbb{R}^{m+n}$. Moreover, it is expected that not only system (2..23) reproduces the evolution of the constraint PDE for marching forward in time, but also it approximates the evolution backward in time of the adjoint model. The aim is to design a numerical solver for the model (2..23) in order to solve constraint PDE and adjoint model with the same methodology and able to deal with non-conservative products.

We are going to prove that in most of the cases the unified system is hyperbolic if constraint PDE as well as the adjoint model, are hyperbolic. In such a case a finite volume scheme is a suitable method for solving this class of problems. Indeed, if (2..1) admits a conservative formulation, there exists $\tilde{\mathbf{K}}(\tilde{\mathbf{U}})$ such that

$$\mathbf{J_R}(\tilde{\mathbf{U}}) = \frac{\partial \tilde{\mathbf{K}}(\tilde{\mathbf{U}})}{\partial \tilde{\mathbf{U}}} \ . \tag{2..25}$$

Then $\mathbf{D} = \mathbf{0}$ and (2..23) is hyperbolic with the same eigenvalues that (2..4). This is stated in the following

PROPOSITION 2..2 If there exists a differentiable function $\tilde{\mathbf{K}}(\tilde{\mathbf{U}})$ such that $\mathbf{J_R}(\tilde{\mathbf{U}}) = \frac{\partial \tilde{\mathbf{K}}(\tilde{\mathbf{U}})}{\partial \tilde{\mathbf{U}}}$, then $\mathbf{D} = \mathbf{0}$ and (2..23) is hyperbolic with eigenvalues

$$\{\lambda_1, \ldots, \lambda_{m+n}\} \ ,$$

each one with multiplicity 2 and eigenvectors

$$\{[\mathbf{v}_1, \mathbf{0}_{m+n}]^T, [\mathbf{0}_{m+n}, \mathbf{v}_1^t]^T, \ldots, [\mathbf{v}_{m+n}, \mathbf{0}_{m+n}]^T, [\mathbf{0}_{m+n}, \mathbf{v}_{m+n}^t]^T\} \ ,$$

where $\mathbf{v}_i$ is the eigenvector of $\mathbf{J_R}$ associated with λ_i and $\mathbf{v}_i^t$ is the eigenvector of $\mathbf{J_R}^T$ associated with λ_i. Here $\mathbf{0}_{m+n}$ is the zero vector of $\mathbb{R}^{m+n}$.

Proof. Let us write system (2..23) in the quasilinear form

$$\partial_t \mathbf{Q} + \mathbf{A}(\mathbf{Q}) \partial_x \mathbf{Q} = \mathbf{S}(\mathbf{Q}) \ , \tag{2..26}$$

where

$$\mathbf{A}(\mathbf{Q}) = \begin{pmatrix} \mathbf{J_R} & \mathbf{0} \\ \mathbf{D} & \mathbf{J_R}^T \end{pmatrix} \ . \tag{2..27}$$

Notice that $det(\mathbf{A} - \mu \mathbf{I}_{2(m+n)}) = det(\mathbf{J_R} - \mu \mathbf{I}_{(m+n)}) det(\mathbf{J_R}^T - \mu \mathbf{I}_{(m+n)}) = p(\mu)^2$, where $p(\mu)$ is the characteristic polynomial of $\mathbf{J_R}$, from which we obtain that eigenvalues of this system are the same of $\mathbf{J_R}$, each one with at lest an algebraic multiplicity 2. It is because it may exists indices i and j such that $i \neq j$ and $\lambda_i = \lambda_j$.

On the other hand, from the existence of $\tilde{\mathbf{K}}$, we note that

$$\frac{\partial \mathbf{J}_{\mathbf{R}i,k}}{\partial \tilde{\mathbf{U}}_j} = \frac{\partial^2 \tilde{\mathbf{K}}(\tilde{\mathbf{U}})}{\partial \tilde{\mathbf{U}}_k \partial \tilde{\mathbf{U}}_j} = \frac{\partial^2 \tilde{\mathbf{K}}(\tilde{\mathbf{U}})}{\partial \tilde{\mathbf{U}}_j \partial \tilde{\mathbf{U}}_k} = \frac{\partial \mathbf{J}_{\mathbf{R}i,j}}{\partial \tilde{\mathbf{U}}_k} \ , \tag{2..28}$$

so from (2..18) we have $\mathbf{D} = \mathbf{0}$. Thus the Jacobian matrix of the system now takes the form

$$\mathbf{A}(\mathbf{Q}) = \begin{bmatrix} \mathbf{J_R} & \mathbf{0} \\ \mathbf{0} & \mathbf{J_R}^T \end{bmatrix} \ . \tag{2..29}$$

In order to find the set of eigenvectors, let us denote by $\mathbf{v}_i$ the eigenvector of $\mathbf{J_R}$ associated with λ_i and $\mathbf{v}_i^t$ by the eigenvector of $\mathbf{J_R}^T$ associated with λ_i. Then is straightforward to verify that

$$\begin{bmatrix} \mathbf{J_R} & \mathbf{0} \\ \mathbf{0} & \mathbf{J_R}^T \end{bmatrix} \begin{bmatrix} \mathbf{v}_i \\ \mathbf{0}_{n+m} \end{bmatrix} = \lambda_i \begin{bmatrix} \mathbf{v}_i \\ \mathbf{0}_{n+m} \end{bmatrix} \tag{2..30}$$

and

$$\begin{bmatrix} \mathbf{J_R} & \mathbf{0} \\ \mathbf{0} & \mathbf{J_R}^T \end{bmatrix} \begin{bmatrix} \mathbf{0}_{n+m} \\ \mathbf{v}_i^t \end{bmatrix} = \lambda_i \begin{bmatrix} \mathbf{0}_{n+m} \\ \mathbf{v}_i^t \end{bmatrix} . \tag{2..31}$$

In this form, a set of eigenvectors is obtained. The verification of the linear independence comes from the fact that $\{\mathbf{v}_1, ..., \mathbf{v}_m\}$ as well as $\{\mathbf{v}_1^t, ..., \mathbf{v}_m^t\}$ are two sets of linearly independent vectors.   □

REMARK 2..2 The previous proposition shows that unified systems coming from conservative hyperbolic equations are always hyperbolic. This does not mean that non-conservative systems are not hyperbolic. So in the case of systems where $\mathbf{D}$ is not the null matrix, the analysis of hyperbolicity has to be done case by case.

2.3. *Existence of solutions, the symetrizable case*

In this section we discuss the setting where a well-posedness theorem for system (2..23) can be formulated. Recall from Proposition 2..2, system (2..23) can be written in the quasilinear form

$$\partial_t \mathbf{Q} + \mathbf{A}(\mathbf{Q}) \partial_x \mathbf{Q} = \mathbf{S}(\mathbf{Q}) , \tag{2..32}$$

with

$$\mathbf{A}(\mathbf{Q}) = \begin{pmatrix} \mathbf{J_R} & \mathbf{0} \\ \mathbf{D} & \mathbf{J_R}^T \end{pmatrix}, \quad \mathbf{S}(\mathbf{Q}) = \begin{pmatrix} \tilde{\mathbf{L}} \\ -\frac{\partial \tilde{\mathbf{L}}}{\partial \tilde{\mathbf{U}}}^T \mathbf{P} + \frac{1}{2} \nabla_{\tilde{\mathbf{U}}} ((\psi - \bar{\psi})^2) \end{pmatrix}. \tag{2..33}$$

In [35], has been made the hypothesis of strict hyperbolicity: that the Jacobian matrix $\mathbf{A}$ has real and distinct eigenvalues. Then, under this assumption, (2..32) is called a symetrizable hyperbolic system provided there exists $\mathbf{A}_0(\mathbf{Q})$ positive-definite, such that $\mathbf{A}_0 \mathbf{A}$ is symmetric. See Proposition 2.2, Chapter 16 in [35] for further details.

The assumption of the symmetry of $\mathbf{A}$ is natural in this context of conservation laws. In [24], Lax introduced a general notion of symmetrizer in the context of pseudodifferential operators to prove that any strictly hyperbolic system is symmetrizable. Also, as was mentioned in [34] under the context of systems of conservation laws having an entropy solution, Godunov [12] and Friedrichs and Lax [11] observed that systems of conservation laws admitting a strongly convex entropy can be symmetrized. It is then under this context of strictly hyperbolic matrices that we state the following theorem for the system (2..32). This theorem is the key point to set a descent method to solve optimal control problems related to conservation laws. Even though its proof is classical, see [35], we outline some key steps for the sake of completeness. Let $I = (-a, b)$ be an interval with $t \in I$.

THEOREM 2..3 Let $\mathbf{Q}(x,0) = f \in H^k(\mathbb{R}^d)$, $k > \frac{d}{2} + 1$. If $\mathbf{A}_0(t,x,\mathbf{Q})$, $\mathbf{A}(t,x,\mathbf{Q})$ in $C(I, H^k \times H^k(\mathbb{R}^d))$ and $\mathbf{L}(\mathbf{Q}, \mathbf{b})$, $\bar{\psi}$, $\psi(\mathbf{Q}, \mathbf{b})$ in $C(I, H^{k+1} \times H^k(\mathbb{R}^d))$, then the initial value problem (2..32) has a unique solution in $C(I, H^k \times H^k(\mathbb{R}^d))$.

Proof. Let $\{J_\varepsilon : 0 < \varepsilon \leqslant 1\}$ a Friedrichs mollifier. The idea is to obtain a solution to (2..32) as the limit of unique solutions $\mathbf{Q}_\varepsilon$ to the systems of ODEs

$$\begin{cases} \mathbf{A}_0(t,x,J_\varepsilon\mathbf{Q}_\varepsilon)\partial_t\mathbf{Q}_\varepsilon + J_\varepsilon\mathbf{A}(t,x,J_\varepsilon\mathbf{Q}_\varepsilon)\partial_x J_\varepsilon\mathbf{Q}_\varepsilon = J_\varepsilon\mathbf{S}(t,x,J_\varepsilon\mathbf{Q}_\varepsilon), \\ \mathbf{Q}_\varepsilon(0) = f, \end{cases} \tag{2..34}$$

for t close to 0. We denote $\mathbf{A}_\varepsilon = \mathbf{A}(t,x,J_\varepsilon\mathbf{Q}_\varepsilon)$.

Let us consider the L^2-inner product

$$(w, \mathbf{A}_{0\varepsilon}w), \quad \mathbf{A}_{0\varepsilon} = \mathbf{A}_0(t,x,J_\varepsilon\mathbf{Q}),$$

which, by the positivity of $\mathbf{A}_0$, defines an equivalent L^2-norm. Therefore by the symmetry of $\mathbf{A}_0$ and (2..34)

$$\frac{d}{dt}(D^\alpha\mathbf{Q}_\varepsilon, \mathbf{A}_{0\varepsilon}(t)D^\alpha\mathbf{Q}_\varepsilon) = -2(D^\alpha(J_\varepsilon\mathbf{A}_\varepsilon\partial_x J_\varepsilon\mathbf{Q}_\varepsilon), D^\alpha\mathbf{Q}_\varepsilon) + 2(D^\alpha J_\varepsilon\mathbf{S}, D^\alpha\mathbf{Q}_\varepsilon)$$
$$-2([D^\alpha, \mathbf{A}_{0\varepsilon}]\partial_t\mathbf{Q}_\varepsilon, D^\alpha\mathbf{Q}_\varepsilon) + (D^\alpha\mathbf{Q}_\varepsilon, \mathbf{A}'_{0\varepsilon}(t)D^\alpha\mathbf{Q}_\varepsilon).$$

Then, using Moser–type [28] estimates and Sobolev embeddings, for $|\alpha| \leqslant k$, one has

$$\frac{d}{dt}(D^\alpha\mathbf{Q}_\varepsilon, \mathbf{A}_{0\varepsilon}(t)D^\alpha\mathbf{Q}_\varepsilon) \leqslant C_k(\|\mathbf{Q}_\varepsilon(t)\|_{C^1})(1 + \|J_\varepsilon\mathbf{Q}_\varepsilon(t)\|^2_{H^k}). \tag{2..35}$$

Since $\|J_\varepsilon\mathbf{Q}_\varepsilon(t)\|_{H^k} \leqslant C\|\mathbf{Q}_\varepsilon(t)\|_{H^k}$, by (2..35), $\mathbf{Q}_\varepsilon$ satisfies

$$\begin{cases} \dfrac{d}{dt}\|\mathbf{Q}_\varepsilon(t)\|^2_{H^k} \leqslant E(\|\mathbf{Q}_\varepsilon(t)\|^2_{H^k}), \\ \|\mathbf{Q}_\varepsilon(0)\|^2_{H^k} = \|f\|^2_{H^k}. \end{cases}$$

Thus by Gronwall's inequality, $\mathbf{Q}_\varepsilon$ exists for t in an interval $[0,b)$ independent of ε and

$$\|\mathbf{Q}_\varepsilon(t)\|_{H^k} \leqslant K(t) < \infty. \tag{2..36}$$

Moreover, by the time-reversibility, $\mathbf{Q}_\varepsilon$ exists and is bounded on $I = (-a,b)$.

By (2..36) and (2..34), $\mathbf{Q}_\varepsilon \in C(I,H^k) \cap C^1(I,H^{k-1})$ is bounded and therefore will have a weak limit point $\mathbf{Q} \in L^\infty(I,H^k(\mathbb{R}^d)) \cap \mathrm{Lip}(I,H^{k-1}(\mathbb{R}^d))$. Moreover, by Ascoli's Theorem there is a sequence, still denoted $\mathbf{Q}_\varepsilon$, such that $\mathbf{Q}_\varepsilon \to \mathbf{Q}$ in $C(I,H^{k-1}(\mathbb{R}^d))$ and since $k > \frac{d}{2}+1$

$$\mathbf{Q}_\varepsilon \to \mathbf{Q} \quad \text{in } C(I,C^1(\mathbb{R}^d)).$$

Consequently, (2..32) follows in the limit from (2..34). Uniqueness is treated in a similar way to the estimates above. This conclude the proof of the Theorem. $\qquad\square$

COROLLARY 2..1 Let $\mathbf{b} \in H^k(\mathbb{R}^d)$, $k > \frac{d}{2}+1$. If $\mathbf{Q} \in C(I,H^k(\mathbb{R}^d))$ is the solution of (2..32) with initial condition $\mathbf{Q}(0) = \mathbf{b}$ then there exists an open neighborhood $V_\mathbf{b}$ of $\mathbf{b}$ in $H^k(\mathbb{R}^d)$ such that system (2..32) with initial condition $\beta \in V_\mathbf{b}$ and right hand side $\mathbf{g} = \partial_\mathbf{Q}\mathbf{S} - \partial_\mathbf{Q}J_R\partial_x\mathbf{Q} \in C(H^k(\mathbb{R}^d))$ have a solution $\mathbf{Q}_\beta$. Moreover the mapping $G : V_\mathbf{b} \to C(H^k(\mathbb{R}^d))$, defined by $G(\beta) = \mathbf{Q}_\beta$, is of class C^∞. Finally, if $V = DG(\beta) \cdot h$, for some $\beta \in V_\mathbf{b}$ and some $h \in H^k(\mathbb{R}^d)$, then V is the unique solution of the problem

$$\partial_t V + J_R(\mathbf{Q}_\beta)\partial_x V = \mathbf{g}V, \quad V(0) = \mathbf{b}.$$

Proof. Proceeding as in the proof of Theorem 2..3, we can consider

$$F : C(H^k(\mathbb{R}^d)) \times H^k(\mathbb{R}^d) \to C(H^k(\mathbb{R}^d)) \times H^k(\mathbb{R}^d)$$

the mapping given by

$$F(\mathbf{Q}, \mathbf{b}) = (\partial_t \mathbf{Q} + \mathbf{A}(\mathbf{Q}) \partial_x \mathbf{Q} - \mathbf{S}(\mathbf{Q}), \mathbf{Q}(0) - \mathbf{b}).$$

Then F is of class C^∞ and

$$\frac{\partial F}{\partial \mathbf{Q}}(\mathbf{Q}, \mathbf{b}) \cdot V = (\partial_t V + J_R(\mathbf{Q}_\beta) \partial_x V - \mathbf{g}V, V(0)).$$

Applying Theorem 2..3 we have that $\frac{\partial F}{\partial \mathbf{Q}}(\mathbf{Q}, \mathbf{b})$ is an isomorphism from $C(H^k(\mathbb{R}^n))$ onto $C(H^k(\mathbb{R}^n)) \times H^k(\mathbb{R}^n)$. Now, if $\mathbf{Q}$ is a solution of (2..32), then $F(\mathbf{Q}, \mathbf{b}) = (0, 0)$. Therefore, by the implicit function theorem, there exists an open neighborhood $V_\mathbf{b}$ of $\mathbf{b}$ in $H^k(\mathbb{R}^d)$ and $G : V_\mathbf{b} \to C(H^k(\mathbb{R}^d))$ such that $F(G(\beta), \beta) = (0, 0)$ for every $\beta \in V_\mathbf{b}$. Moreover, G is also of class C^∞ and

$$\frac{\partial F}{\partial \mathbf{Q}}(\mathbf{Q}_\beta, \beta) \circ DG(\beta) \cdot h + \frac{\partial F}{\partial \beta}(\mathbf{Q}_\beta, \beta) \cdot h = (0, 0) \quad \forall h \in H^k(\mathbb{R}^d).$$

Then, if we set $V = DG(\beta) \cdot h$, we get

$$\partial_t V + J_R(\mathbf{Q}_\beta) \partial_x V = \mathbf{g}V, \quad V(0) = \mathbf{b}.$$

$\square$

Note that we have established the results above for $x \in \mathbb{R}^d$ and $k > \frac{d}{2} + 1$. This is to keep the generality of this results. In this work will be enough to consider $d = 1$ and $k > 3/2$ as we did in the formulation of system (2..1).

2.4. *The optimal control problem*

One of our main motivations applying an optimal control strategy to general conservation laws is the possibility of study numerical inverse problems related to this type of systems. Example of them are the bottom detection in water-waves type system, see [29], or the tracking of initial conditions in the presence of shocks [5, 25], to mention but a few.

We are considering the functional

$$J(\mathbf{Q}, \mathbf{b}) = \frac{1}{2} \int_0^T \int_\mathbb{R} |\psi(\mathbf{Q}, \mathbf{b}) - \bar{\psi}|^2 dx dt + \frac{\alpha}{2} \|\mathbf{b}\|_{H^k}^2,$$

with $\bar{\psi}(t, x) \in C((0, T), H^k(\mathbb{R}^d))$, $k > \frac{d}{2} + 1$, a given function and $\alpha > 0$. We want to find $\mathbf{b}$ such that J is small. The nature of ψ is not specified and it depends on the physical problem under consideration, as well as the itself election of J. The admissible control set is a nonempty convex closed subset U_{ad} of $H^k(\mathbb{R}^d)$.

Then the optimal control problem is formulated in the following way

$$\left. \begin{array}{c} \text{Minimize } J(\mathbf{Q}, \mathbf{b}), \\ (\mathbf{Q}, \mathbf{b}) \in H^1(H^k(\mathbb{R}^d)) \times U_{ad} \text{ verifies (2..32).} \end{array} \right\} \tag{P}$$

Now we study the existence of a solution of (P).

THEOREM 2..4 Let $k > \frac{d}{2} + 1$. If there exists a feasible pair $(\mathbf{Q}, \mathbf{b}) \in U_{ad} \times H^1(H^k(\mathbb{R}^d))$ satisfying (2..32), then there exists at least one optimal solution $(\mathbf{Q}_0, \mathbf{b}_0)$ of (P).

Proof. Since there is one feasible pair for the problem and J is bounded below by zero we may take a minimizing sequence $\{(\mathbf{Q}_n, \mathbf{b}_n)\} \subset U_{ad} \times H^1(H^k(\mathbb{R}^d))$. Then $\frac{\alpha}{2} \|\mathbf{b}_n\|_{H^k}^2 \leqslant J(\mathbf{Q}_n, \mathbf{b}_n) < \infty$, which implies that $\{\mathbf{b}_n\}$ is a bounded sequence in $H^k(\mathbb{R}^d)$. Thus, we can take a subsequence, denoted in the same way, such that $\mathbf{b}_n \to \mathbf{b}_0$ weakly in $H^k(\mathbb{R}^d)$. Since U_{ad} is closed and convex, $\mathbf{b}_0 \in U_{ad}$. On the other hand, following the proof of Theorem 2..3, we have that $\{\mathbf{Q}_n\}$ is bounded in $H^1(H^k(\mathbb{R}^d))$. Therefore we can assume, by taking a subsequence if necessary, that $\mathbf{Q}_n \to \mathbf{Q}_0$ weakly in $H^1(H^k(\mathbb{R}^d))$. Using the continuity of the inclusion $H^1(H^k(\mathbb{R}^d)) \subset C(H^{k-1}(\mathbb{R}^d))$ we can pass to the limit in system (2..32) satisfied by $(\mathbf{Q}_n, \mathbf{b}_n)$ and to conclude that $(\mathbf{Q}_0, \mathbf{Q}_0)$ also satisfies (2..32). Namely, $(\mathbf{Q}_0, \mathbf{b}_0)$ is a feasible pair for problem (P). Taking into consideration that $J(\mathbf{Q}, \mathbf{b})$ is weakly lower semi-continuous, the result follows. $\qquad \square$

Let us remark that the assumption of Theorem 2..4 can be checked by taking $\mathbf{Q} \in H^1(H^k(\mathbb{R}^d))$ with $\mathbf{Q}(0) = f$. This uniquely defines $\mathbf{b}$ from the partial differential equation (2..32) as $\mathbf{b} = \mathbf{Q}(0)$. If $\mathbf{b} \in U_{ad}$ then the assumption is satisfied. This is the case if $U_{ad} = H^k(\mathbb{R}^d)$.

Finally we state the following proposition giving the optimality conditions satisfied by the solutions $(\mathbf{Q}_0, \mathbf{b}_0)$ of (P). Its proof is based on Corollary 2..1 and the results obtained in Section 2.1..

PROPOSITION 2..5 Let $k > \frac{d}{2} + 1$ and assume that $(\mathbf{Q}_0, \mathbf{b}_0)$ is a solution of (P). Then there exist a unique element $\mathbf{P}_0 \in C(H^k(\mathbb{R}^d))$ such that the following system is satisfied

$$\partial_t \mathbf{Q}_0 + \mathbf{J}_R \partial_x \mathbf{Q}_0 = \mathbf{L}(\mathbf{Q}_0) \,,$$
$$\mathbf{Q}_0(x, T) = \mathbf{b}_0 \,,$$

$$\partial_t \mathbf{P}_0 + \mathbf{J}_R^T \partial_x \mathbf{P}_0 = \frac{\partial \mathbf{L}}{\partial \mathbf{Q}_0}^T \mathbf{P}_0 + \frac{1}{2} \nabla_{\mathbf{Q}_0}((\psi - \bar{\psi})^2) - \mathbf{D} \partial_x \mathbf{Q}_0 \,,$$
$$\mathbf{P}_0(x, T) = \mathbf{0} \,,$$

$$\mathbf{P}_0|_{t=\mathbf{0}}.$$

3. The numerical scheme

In this section, we present the scheme for solving simultaneously conservative as well as non-conservative hyperbolic balance laws of the form (2..23). Cell averages are evolved by following the one-step finite volume formula

$$\mathbf{Q}_i^{n+1} \quad = \quad \mathbf{Q}_i^n - \eta \frac{\Delta t}{\Delta x}(\mathbf{F}_{i+\frac{1}{2}} - \mathbf{F}_{i-\frac{1}{2}}) - \eta \frac{\Delta t}{\Delta x}(\mathbf{D}_{i+\frac{1}{2}} + \mathbf{D}_{i-\frac{1}{2}}) \\ - \frac{\eta}{\Delta x}(\mathbf{B}\partial \mathbf{Q})_i + \eta \Delta t \mathbf{S}_i \,, \tag{3..1}$$

where η controls the evolution of the system. This is because in the context of the adjoint method we need the constraint PDE evolves forward in time up to T and then the adjoint model evolves backward in time from T to 0. So, $\eta = 1$ makes the forward in time and $\eta = -1$ makes the backward in time. This formula is a conventional one, available in the literature for solving non-conservative schemes, see for instance [10]. However, we are going to modify the usual implementation of these type of schemes in order to account for the correct wave propagation. By the sake of completeness, we provide a detailed

description of the scheme. We have

$$(\mathbf{B}\partial\mathbf{Q})_i = \frac{1}{\Delta t \Delta x} \int_t^{t^{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{B}(\mathbf{Q}_i(\xi,\tau))\partial_x\mathbf{Q}_i(\xi,\tau)d\xi d\tau\,,$$

$$\mathbf{S}_i = \frac{1}{\Delta t \Delta x} \int_t^{t^{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{S}(\mathbf{Q}_i(\xi,\tau))d\xi d\tau \tag{3..2}$$

and

$$\mathbf{F}_{i+\frac{1}{2}} = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \tilde{\mathbf{F}}_h(\mathbf{Q}_{i+\frac{1}{2}}^-(\tau),\mathbf{Q}_{i+\frac{1}{2}}^+(\tau))d\tau\,,$$

$$\mathbf{D}_{i+\frac{1}{2}} = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \tilde{\mathbf{D}}(\mathbf{Q}_{i+\frac{1}{2}}^-(\tau),\mathbf{Q}_{i+\frac{1}{2}}^+(\tau))d\tau\,, \tag{3..3}$$

where

$$\tilde{\mathbf{F}}_h(\mathbf{Q}_h^-(\tau),\mathbf{Q}_h^+(\tau)) =$$
$$\frac{1}{2}(\mathbf{F}(\mathbf{Q}_h^-)+(\mathbf{F}(\mathbf{Q}_h^+))) - \eta\frac{1}{2}\int_0^1 |\mathbf{A}(\Psi(s;\mathbf{Q}_h^-,\mathbf{Q}_h^+))|\frac{d\Psi}{ds}ds\,. \tag{3..4}$$

Moreover $\mathbf{A} := \frac{\partial\mathbf{F}}{\partial\mathbf{Q}}$, with $|\mathbf{A}| = \mathbf{R}|\Lambda|\mathbf{R}^{-1}$, where $|\Lambda| = diag(|\lambda_1|,...,|\lambda_m|)$ and then numerical jumps are obtained as

$$\tilde{\mathbf{D}}(\mathbf{Q}_h^-(\tau),\mathbf{Q}_h^+(\tau)) = \eta \cdot \frac{1}{2}\int_0^1 \mathbf{B}(\Psi(s;\mathbf{Q}_h^-,\mathbf{Q}_h^+)\frac{d\Psi}{ds}ds\,, \tag{3..5}$$

with

$$\Psi(s;\mathbf{Q}_h^-,\mathbf{Q}_h^+) := \mathbf{Q}_h^- + s(\mathbf{Q}_h^+ - \mathbf{Q}_h^-)\,. \tag{3..6}$$

Equation (3..4) corresponds to the DOT (Dumbser-Osher-Toro) Riemann solver, introduced in [9]. The integrals are evaluated numerically, by using some quadrature rule. Here $\mathbf{Q}_i(\xi,\tau)$ is a predictor inside the computational cell $[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}]$ and $\mathbf{Q}_{i+\frac{1}{2}}^\pm(\tau)$ are extrapolations of the solution at both sides of the cell interface position $x = x_{i+\frac{1}{2}}$. They are usually computed by using the so-called Cauchy-Kowalewsky procedure and Taylor series expansions.

3.1. *The predictor step for second order of accuracy*

In this section, we deal with the strategy to get predictor inside the computational cell. This stage requires two ingredients, the first one is a polynomial representation of the solution and then a local evolution in time of approximate values within the computational cell at located spatial positions. The second stage is given by the evolution of values within the computational cell, the resulted state of the local evolution is commonly referred to as the predictor.

The polynomial representation of the solution is carried in terms of the so called *reconstruction procedure*, [14]. Let us define the reconstruction polynomial within cell $[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}]$ as

$$\mathbf{P}_i(\xi) = \mathbf{Q}_i^n + (\xi - \frac{1}{2})\Delta_i\,, \tag{3..7}$$

with Δ_i being a slope endowing the scheme with the Total Variation Diminishing property. In this work we use the so-called Minmod limiter, see the pioneering work [36] about TVD flux limiters for hyperbolic conservation laws.

In a component wise, it has the form

$$\Delta_{i,j} = \begin{cases} 0 & , \quad if \; (\mathbf{Q}_{i,j}^n - \mathbf{Q}_{i-1,j}^n) \cdot (\mathbf{Q}_{i+1,j}^n - \mathbf{Q}_{i,j}^n) \leqslant 0 \,, \\ (\mathbf{Q}_{i,j}^n - \mathbf{Q}_{i-1,j}^n) & , \quad if \; |\mathbf{Q}_{i,j}^n - \mathbf{Q}_{i-1,j}^n| < |(\mathbf{Q}_{i+1,j}^n - \mathbf{Q}_{i,j}^n)| \,, \\ (\mathbf{Q}_{i+1,j}^n - \mathbf{Q}_{i,j}^n) & , \quad if \; |\mathbf{Q}_{i+1,j}^n - \mathbf{Q}_{i,j}^n| < |(\mathbf{Q}_{i,j}^n - \mathbf{Q}_{i-1,j}^n)| \,. \end{cases} \tag{3..8}$$

Polynomial $\mathbf{P}_i$ is defined in terms of a local variable $\xi \in [0,1]$ and it is related with the computational cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ through out the relationship $x = x_{i-\frac{1}{2}} + \xi \Delta x$.

The predictor is an approximation of the solution within the cell, which represents the local evolution of the solution at x. This evolution in time is given in terms of a Taylor series expansion

$$\mathbf{Q}_i(\xi,\tau) = \mathbf{Q}(\xi,0_+) + \tau \partial_t \mathbf{Q}(\xi,0_+) \,. \tag{3..9}$$

Notice that the series is truncated up to the first spatial derivative. This is enough to produce solutions of second order of accuracy. The time derivative is expressed in terms of purely spatial derivatives. This is called the Cauchy-Kowalewsky procedure and it is based on the use of the governing equation

$$\tau \partial_t \mathbf{Q}(\xi,\tau) = -\mathbf{A}(\mathbf{Q}(\xi,\tau)) \partial_x \mathbf{Q}(\xi,\tau) \,. \tag{3..10}$$

Replacing into the expansion, we have

$$\mathbf{Q}_i(\xi,\tau) = \mathbf{Q}(\xi,0_+) - \tau \mathbf{A}(\mathbf{Q}(\xi,0_+)) \partial_x \mathbf{Q}(\xi,0_+) \,. \tag{3..11}$$

Notice that to obtain the evolution at time τ, we require an approximation of the solution at local time $\tau = 0$, which is associated with the global time level t^n. As finite volume only provide an approximation of cell averages, the punctual information within cells is not straightforward obtained. In this stage we use the polynomial representation $\mathbf{P}_i(\xi)$ of the solution at time level t^n, $\mathbf{Q}_i^n$, so the Taylor series expansion takes the form

$$\mathbf{Q}_i(\xi,\tau) = \mathbf{P}_i(\xi) - \tau \mathbf{A}(\mathbf{P}(\xi)) \mathbf{P}_i'(\xi) \frac{1}{\Delta x} \,, \tag{3..12}$$

which is a polynomial in time. Using this polynomial, we evaluate the integrals (3..2) and (3..3). In particular, for the flux evaluation, we require

$$\begin{array}{rcl} \mathbf{Q}_{i+\frac{1}{2}}^-(\tau) & = & \mathbf{Q}_{i+1}(0,\tau) \,, \\ \mathbf{Q}_{i+\frac{1}{2}}^+(\tau) & = & \mathbf{Q}_i(1,\tau) \,. \end{array} \tag{3..13}$$

3.2. *Treatment for forward and backward evolution in time*

The proposed method uses approximations of classical Riemann problem at cell interfaces, for computing the integrals in (3..3). This is carried out for both the forward ($\eta = 1$) and backward ($\eta = -1$) evolution in time, as hyperbolic equations are not reversible in general we may obtain different results and then we must select the right value for the minimization problem. To illustrate this fact, let us consider the Burgers equation in the context of the unified system. So, the governing equation associated to the constraint PDE, has the form

$$\partial_t q + \eta q \partial_x q = 0 \,, \tag{3..14}$$

where $\eta = 1$ for forward evolution in time and $\eta = -1$ for backward evolution in time. If we study Riemann problems associated to the unified system but restricted to the state variable, we have

$$\partial_t q + \eta q \partial_x q = 0 , t > 0 ,$$
$$q(x,0) = \begin{cases} q_L , x < 0 , \\ q_R , x \geqslant 0 , \end{cases} \tag{3..15}$$

with q_L and q_R constant states. The Riemann problem solution of this problem is well studied in the literature, see for instance [36]. Let us assume that $q_L > 0 > q_R$ and $\eta = 1$, so the solution at the interface position is a shock wave propagating with velocity $S = \frac{q_L + q_R}{2} < q_L$. Therefore, at the interface position it has the value $q(0,t) = q_L$. If now, we solve backward in time the unified problem, we set $\eta = -1$, so we have

$$\partial_t q - q \partial_x q = 0 ,$$
$$q(x,0) = \begin{cases} q_L , x < 0 , \\ q_R , x \geqslant 0 . \end{cases} \tag{3..16}$$

Then, by using the change of variable $x = -x$, we obtain

$$\partial_t q + q \partial_x q = 0 ,$$
$$q(x,0) = \begin{cases} q_R , x < 0 , \\ q_L , x \geqslant 0 , \end{cases} \tag{3..17}$$

and so the solution corresponds to a rarefaction wave and thus, at the interface position, takes the value $q(0,t) = 0$.

On the other hand, the adjoint state in the backward evolution involves q but it requires the value generated in the forward evolution, in this particular case the shock wave and not the right value of q generated in the backward evolution (a rarefaction wave). If we choose the rarefaction wave, then we sub estimate the required values and thus it may have an impact in the minimization procedure.

To force the adjoint state to take the wave propagation of interest for us, we record the state variables from the solution obtained in the forward evolution in time and then use it to provide the evolution of the adjoint state in the backward evolution. That means for $\eta = -1$ we only use (3..1) to evolve the variables associated to the adjoint states and we froze the evolution of state variables. The suggested application of the strategy is listed below.

1) Given $\mathbf{b}^k$, we solve the constraint PDE, using the one-step finite volume formula (3..1) with $\eta = 1$ up to the output time $t = T$ and record from $\mathbf{Q}$ the variables associated to the state variables $\mathbf{U}$. Final time T is reached in a finite number of steps, let say n_T. That means, starting from $t^0 = 0$, we do $t^{n+1} = t^n + \Delta t$, n_T times.

2) Solve the adjoint problem. We use the same formula (3..1) with $\eta = -1$ and the initial condition $\mathbf{Q}^{n_T}_{j+m+n} = \mathbf{P}^n_j = 0$, but applied in a different way as in step 1). We turn off the evolution associated with the state variables. That means, we do

$$\mathbf{P}^n_{i,j} = \mathbf{P}^{n+1}_{i,j} + \frac{\Delta t}{\Delta x} \left(\mathbf{F}_{i+\frac{1}{2},j+m+n} - \mathbf{F}_{i-\frac{1}{2},j+m+n} \right)$$
$$+ \frac{\Delta t}{\Delta x} \left(\mathbf{D}_{i+\frac{1}{2},j+m+n} + \mathbf{D}_{i-\frac{1}{2},j+m+n} \right) + \frac{1}{\Delta x} (\mathbf{B}\partial\mathbf{Q})_{i,j+m+n} - \Delta t \mathbf{S}_{i,j+m+n} , \tag{3..18}$$

for $j = 1, ..., (m+n)$. Subindex $j+m+n$ in $\mathbf{F}_{i+\frac{1}{2}}$, $\mathbf{D}_{i+\frac{1}{2}}$, $(\mathbf{B}\partial\mathbf{Q})_i$ and $\mathbf{S}_i$ represent the $(j+m+n)th$ component. For computing the integrals (3..2) and (3..3) we require also the state variables, but

those are used from the record carried out in step 1). Index $n+1$ is consistent with the backward evolution in time, that means starting from $t^0 = T$, we do $t^{n+1} = t^n - \Delta t$. So in a finite number of steps, n_T, we reach $t^{n_T} = 0$.

3) Compute the gradient of the cost function (2..2) using variables associated to the adjoint variables as well as state variables. In this paper as shown formally in (2..20), they normally have the form $\nabla J_i = \mathbf{P}_i^{n_T}$ (expressed in a discrete form which is normally used in simulations). Notice that this procedure can have more general structures and may also depend on the state variables, so let us express the gradient formally in terms of a functional, $G(\mathbf{U}, \mathbf{P}, \mathbf{b})$. So the update of the sought parameters has the form

$$\mathbf{b}_i^{k+1} = \mathbf{b}_i^k - \lambda_{IP} \cdot G(\mathbf{U}, \mathbf{P}, \mathbf{b}^k)_i \,, \tag{3..19}$$

where λ_{IP} is a prescribed constant value. Notice that, the procedure is carried out for finding parameters. However, for another type of problems, for example for finding initial conditions, the procedure is quite similar to the present one.

4) Stop the procedure using some stop criterion. This is normally carried in terms of the relative error between two successive approximations.

4. Numerical examples

In this section, we solve PDE-constrained optimization problems using the numerical scheme described in section 3.. In order to assess the performance of the present scheme, we compare with a conventional solver and so we mimic the strategy usually employed for solving this type of problems, see the reference scheme in Appendix A.

Simulations carried out in this section are obtained with the stable time step, Δt, computed as

$$\Delta t = C_{cfl} \tfrac{\Delta x}{\lambda_\infty} \,, \tag{4..1}$$

where

$$\lambda_\infty = \max_i \max_{j=1,\ldots,2(m+n)} |\lambda_j(\mathbf{H}(x_i))| \,, \tag{4..2}$$

with λ_j being the jth eigenvalue of the Jacobian matrix and $\mathbf{H}(x_i)$ the initial condition evaluated at the barycentre of cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, we choose $C_{cfl} = 0.1$. Notice that, if the initial condition $\mathbf{H}(x)$ changes as a consequence of the change in the sought parameters, the time step must also change accordingly to (4..1).

4.1. *The Burgers equation*

Let us consider the optimal problem consisting of finding the initial condition $h_0(x)$ such that the solution of the following hyperbolic conservation law

$$\partial_t q + \partial_x(\tfrac{q^2}{2}) = 0 \,, t \in [0, T] \,,$$
$$q(x, 0) = h_0(x) \,, \tag{4..3}$$

minimizes the functional

$$J(h_0) = \frac{1}{2} \int_0^T \int_{-\infty}^\infty (q(x,t) - \bar{q}(x,t))^2 dxdt \,, \tag{4..4}$$

where $\bar{q}$ is some measurement chosen to be the exact solution of the following Burgers equation at $t = T$

$$\partial_t q + \partial_x\left(\frac{q^2}{2}\right) = 0 \ , t \in [0, T] \ ,$$
$$q(x, 0) = \bar{h}_0(x) \ ,$$

(4..5)

where $\bar{h}_0(x)$ is a prescribed function. These problems are endowed with transmissive boundary conditions. The adjoint problem is given by

$$\partial_t p + q \partial_x p = \bar{q} - q \ , t \leqslant T \ ,$$
$$p(x, T) = 0 \ .$$

(4..6)

See [25] for further details concerning the derivation of the adjoint problem. The gradient of the cost function is

$$\nabla J(h_0) = p(x, 0) \ .$$

(4..7)

The unified formulation, as proposed in this paper takes the form

$$\partial_t \mathbf{Q} + \partial_x \mathbf{F}(\mathbf{Q}) + \mathbf{B}(\mathbf{Q}) \partial_x \mathbf{Q} = \mathbf{S}(\mathbf{Q}) \ ,$$

(4..8)

with

$$\mathbf{Q} = \begin{bmatrix} q \\ p \end{bmatrix}, \ \mathbf{F}(\mathbf{Q}) = \begin{bmatrix} \frac{q^2}{2} \\ 0 \end{bmatrix}, \ \mathbf{S}(\mathbf{Q}) = \begin{bmatrix} 0 \\ \bar{q} - q \end{bmatrix},$$

(4..9)

and

$$\mathbf{B}(\mathbf{Q}) = \begin{bmatrix} 0 & 0 \\ 0 & q \end{bmatrix} \ .$$

(4..10)

System (4..8) is solved from $t = 0$ up to $t = T$, thus we need an initial condition at $t = 0$. Notice that $p(x, t)$ has a kind of initial condition only at $t = T$. At $t = 0$ we have to set information for $p(x, t)$. However, as p does not influences the evolution of q we set $p(x, 0) = p_0 = 0$. So, the initial condition is $\mathbf{Q}(x, 0) = [h_0(x), 0]^T$. On the other hand, at $t = T$, the second component is set to zero which means

$$\mathbf{Q}(x, T) = [q(x, T), 0]^T \ .$$

In this way, the evolution of the adjoint variable begins with the right initial condition as required in (4..6). We are going to implement the procedure for continuous as well as discontinuous initial conditions.

In order to assess the performance of the solver we compute the error given by $Error = max_i |h_{0,i} - \bar{h}_0(x_i)|$, where $h_{0,i}$ is the approximated initial condition at the cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$.

4.1..1 The Burgers equation with discontinuous initial condition

Let us consider the case in which the aim is to find a discontinuous initial condition. So we setup the optimization problem as follows. We set a starting guess for the procedure, in this case we set
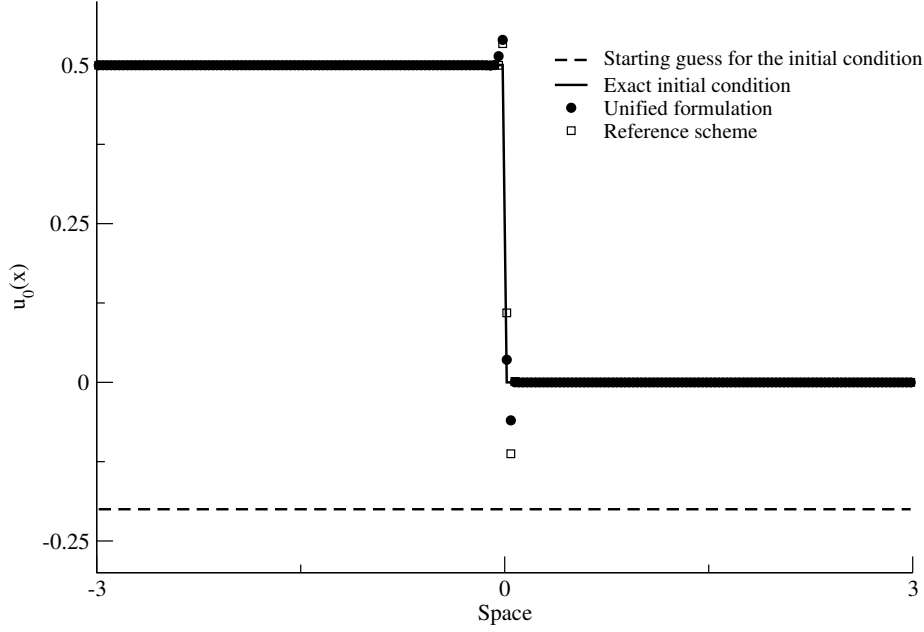
FIG. 1. Burgers, discontinuous initial condition. Sought initial condition, results of unified (circles) and conventional (squares) formulations. The constraint PDE problem is evolved up to $T = 0.12$.

$h_0(x) = -0.2$. We solve both constraint PDE and the adjoint problem on the computational domain $[-3,3]$ using 200 uniform cells. The expected initial condition is

$$\bar{h}_0(x) = \left\{ \begin{array}{l} \frac{1}{2}, x < 0, \\ 0, x \geqslant 0. \end{array} \right. \tag{4..11}$$

To update the sought initial condition we use (3..19) with $\lambda_{IP} = 0.7$ and $G(\mathbf{U}, \mathbf{P}, \mathbf{b}^k)$ given by (4..7). We also implement the reference scheme described in appendix A and using the same information described above.

The results are depicted in Figure 1 at the 80th iteration. The unified formulation has $Error = 6.05 \cdot 10^{-2}$ whereas the solution with the reference solution called here as the conventional solver has $Error = 0.11$. Figure 2 depicts the error at each iteration of the global procedure. We observe that the error is not so good for both methods even when the unified formulation provides the best approximation, it shows undershoot and overshoot at the interface position but the undershoot is less than that resulting from the conventional solver.

Note that in this test, the interface position is found and also the mean shape of the initial profile. However, we observe that the undershot and also the overshoot appear near to the 18th iteration as is shown in Figure 2. Thus it propagates for a while and then it is stabilized. However, these undershoot and overshoot do not disappear.
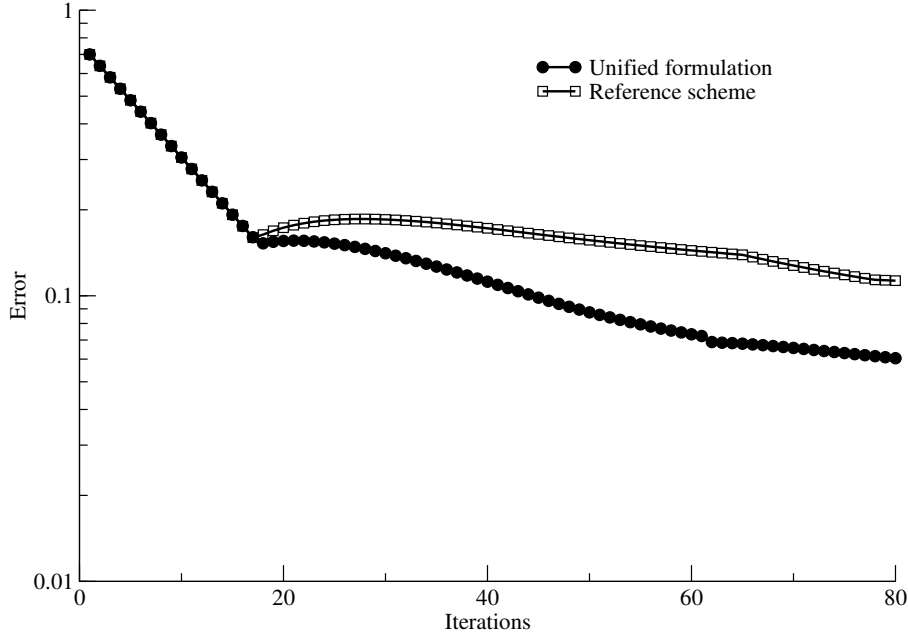
FIG. 2. Burgers, discontinuous initial condition. Comparison between errors associated with the unified formulation (squares) and conventional formulation (circles). The constraint PDE problem is evolved up to $T = 0.12$.

4.1..2 The Burgers equation with continuous initial condition

Let us consider the case in which the aim is to find a continuous initial condition. So we setup the descent algorithm as follows. The optimization procedure is initialized with $h_0(x) = 0$. We solve both constraint PDE and the adjoint problem on the computational domain $[0,1]$ using 160 uniform cells. The expected initial condition is

$$\bar{h}_0(x) = sin(2\pi x) \ . \tag{4..12}$$

To update the sought initial condition we use (3..19) with $\lambda_{IP} = 2.7$ and $G(\mathbf{U}, \mathbf{P}, \mathbf{b}^k)$ given by (4..7). We also implement the reference scheme described in appendix A.

The results are depicted in Figure 3 at the 40th iteration. The unified formulation has $Error = 6.68 \cdot 10^{-3}$ whereas the solution with the reference scheme has $Error = 1.08 \cdot 10^{-2}$. Figure 4 depicts the error at each iteration of the global procedure. We observe that the error for the present unified strategy is one-order of magnitude more accurate than the reference scheme. If the variant described in section 3.2. is not included, that means conventional finite volume implementation, the numerical scheme presented in section 3.) is of second order of accuracy. It seems to be that the high accuracy is still present even if the state variable is frozen in the evolution backward in time associated with the adjoint problem. A CPU time measurement gives us that the present scheme is almost three times more expensive than the reference scheme. However, the solution of the constraint PDE, adjoint model and approximation update, take about three seconds, (it depends on Δt), even so the scheme is very efficient in terms of computational cost.
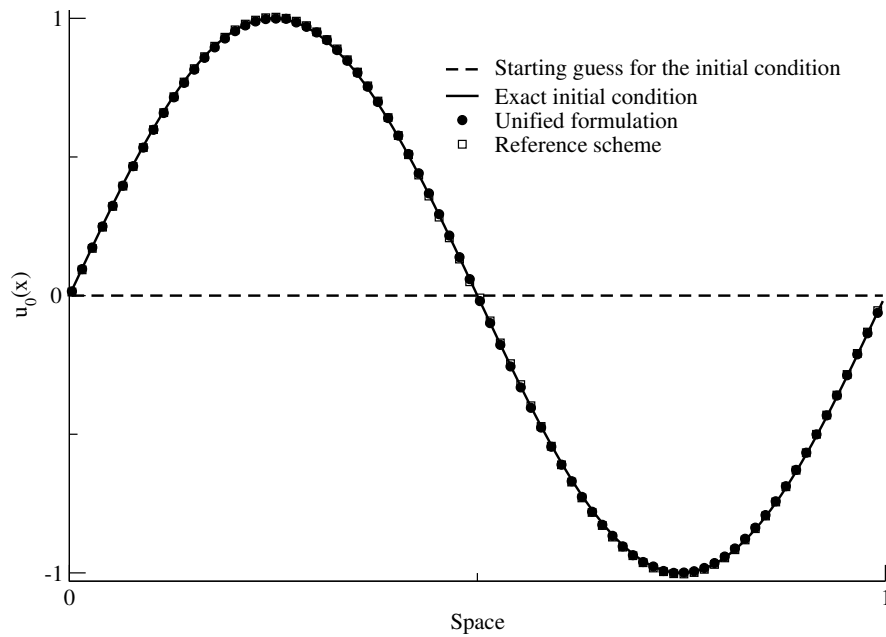
FIG. 3. Burgers, smooth initial condition. Sought initial condition, results of unified (circles) and conventional (squares) formulations. The constraint PDE problem is evolved up to $T = 0.1$.
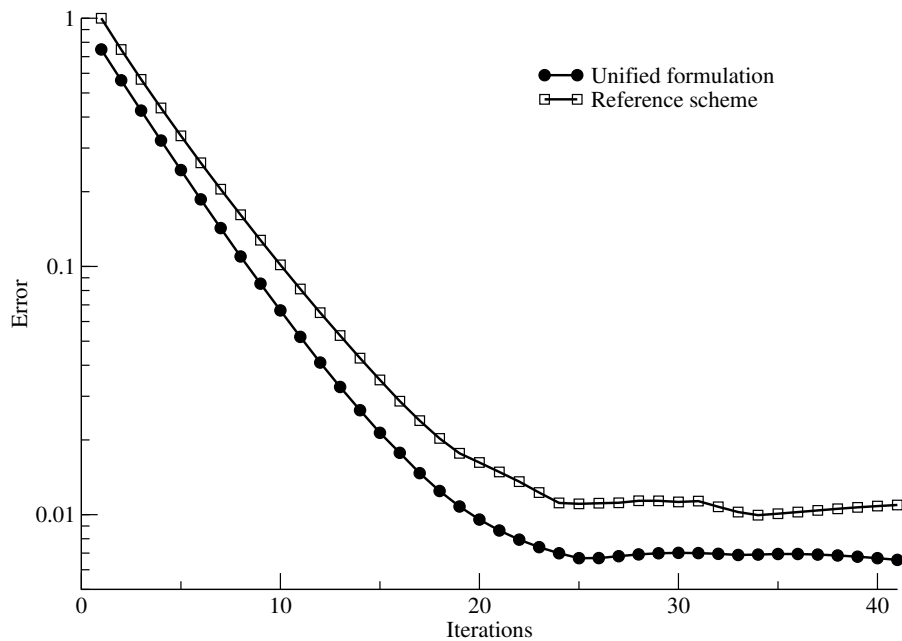

FIG. 4. Burgers, smooth initial condition. Comparison between errors associated with the unified formulation (squares) and conventional formulation (circles). The constraint PDE problem is evolved up to $T = 0.1$.

4.2. *The shallow water equations*

In this section we deal with the solution of the shallow water equation written as follows

$$\partial_t \begin{bmatrix} h \\ q \end{bmatrix} + \partial_x \begin{bmatrix} \varepsilon q \\ \varepsilon \frac{q^2}{h} + \frac{1}{2} \frac{h^2}{\varepsilon} \end{bmatrix} = \begin{bmatrix} 0 \\ -h \partial_x b(x) \end{bmatrix} , \tag{4..13}$$

where $\zeta(x,t)$ and $b(x)$ are the free surface and bottom parameterization respectively, $h(x,t) = 1 + \varepsilon(\zeta - b)$ is the total depth, $u(x,t)$ is the total depth averaged velocity, ε is a dimensionless parameter and $q = hu$ [23].

There is an extensive literature concerning the general water-waves equations and shallow water approximations regimes [19, 22, 23]. Among the many applications to coastal engineering we focus on the bottom detection through measurements on the free surface ζ. This problem has practical interests [30] as well as theoretical challenges [2]. Because of the difficulties arising in the general water-wave system, many asymptotic models are proposed to study this phenomenon [4, 16, 23, 33].

We are going to consider the nonlinear shallow water equations (or Saint-Venant) [8] for study here the inverse problem consisting of finding the bottom surface $b(x)$ such that a prescribed free surface $\bar{\zeta}(x,t)$ is achieved. The adjoint problem associated to this problem is

$$\partial_t \begin{bmatrix} \tilde{h} \\ \tilde{q} \end{bmatrix} + \begin{bmatrix} 0 & \frac{h}{\varepsilon} - \varepsilon \frac{q^2}{h^2} \\ \varepsilon & 2\varepsilon \frac{q}{h} \end{bmatrix} \partial_x \begin{bmatrix} \tilde{h} \\ \tilde{q} \end{bmatrix} = \begin{bmatrix} \frac{\bar{\zeta} - \zeta}{\varepsilon} + \tilde{q} \partial_x b \\ 0 \end{bmatrix} , \tag{4..14}$$

and the gradient of this functional is given by

$$\nabla J = \frac{h-1}{\varepsilon} + b - \bar{\zeta} + (h\tilde{q}) . \tag{4..15}$$

We neglect the term $(h\tilde{q})$ and take $\lambda_{IP} \neq 1$. In this way the scheme is forced to do more than one iteration and the computed free surface has to be very close to the expected one.

Notice that, the bottom surface is a prescribed function, with $\partial_t \mathbf{b}(x) = 0$ so it can be included into the model as a non-evolutionary variable, it would allow us to include non-smooth bottom surfaces. So, the model now takes the form

$$\partial_t \begin{bmatrix} h \\ q \\ b \end{bmatrix} + \partial_x \begin{bmatrix} \varepsilon q \\ \varepsilon \frac{q^2}{h} + \frac{1}{2} \frac{h^2}{\varepsilon} \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & h \\ 0 & 0 & 0 \end{bmatrix} \partial_x \begin{bmatrix} h \\ q \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} , \tag{4..16}$$

The unified formulation for this test is given by

$$\partial_t \mathbf{Q} + \partial_x \mathbf{F}(\mathbf{Q}) + \mathbf{B}(\mathbf{Q}) \partial_x \mathbf{Q} = \mathbf{S}(\mathbf{Q}) , \tag{4..17}$$

where

$$
\mathbf{Q} = \begin{bmatrix} h \\ q \\ b \\ \tilde{h} \\ \tilde{q} \end{bmatrix}, \mathbf{F(Q)} = \begin{bmatrix} \varepsilon q \\ \varepsilon \frac{q^2}{h} + \frac{1}{2}\frac{h^2}{\varepsilon} \\ 0 \\ 0 \\ 0 \end{bmatrix},
$$

$$
\mathbf{B(Q)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & h & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{h}{\varepsilon} - \varepsilon\frac{q^2}{h^2} \\ 0 & 0 & -\tilde{q} & \varepsilon & 2\varepsilon\frac{q}{h} \end{bmatrix}, \mathbf{S(Q)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \frac{\zeta - \zeta}{\varepsilon} \end{bmatrix}.
$$

$$(4..18)$$

Notice that the Jacobian matrix of $\mathbf{F}$ with respect to $\mathbf{Q}$, $\mathbf{A(Q)}$ is given by

$$
\mathbf{A(Q)} = \begin{bmatrix} 0 & \varepsilon & 0 & 0 & 0 \\ \frac{h}{\varepsilon} - \frac{q^2}{h^2}\varepsilon & 2\varepsilon\frac{q}{h} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{4..19}
$$

Moreover, this has the decomposition $\mathbf{A = R\Lambda R^{-1}}$, with

$$
\mathbf{R(Q)} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ \frac{q^2\varepsilon^2 - h^3}{hq\varepsilon^2 + h^{\frac{5}{2}}\varepsilon} & \frac{q^2\varepsilon^2 - h^3}{hq\varepsilon^2 - h^{\frac{5}{2}}\varepsilon} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{4..20}
$$

$$
\mathbf{\Lambda(Q)} = \begin{bmatrix} u\varepsilon - \sqrt{h} & 0 & 0 & 0 & 0 \\ 0 & u\varepsilon + \sqrt{h} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{4..21}
$$

$$
\mathbf{R(Q)}^{-1} = \begin{bmatrix} \frac{q^2\varepsilon^2 - h^3}{2h^{\frac{3}{2}}q\varepsilon - 2h^3} & -\frac{\varepsilon}{2\sqrt{h}} & 0 & 0 & 0 \\ -\frac{q^2\varepsilon^2 - h^3}{2h^{\frac{3}{2}}q\varepsilon + 2h^3} & \frac{\varepsilon}{2\sqrt{h}} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{4..22}
$$

In this test, the aim is to find the bottom surface $b(x)$ such that the model system (4..13) provides a free surface $\zeta(x,t)$ close to the profile $\bar{\zeta}(x,t) = \frac{0.3(x-t)}{cosh(x-t)^2}$. Here (4..13) is endowed with transmissive
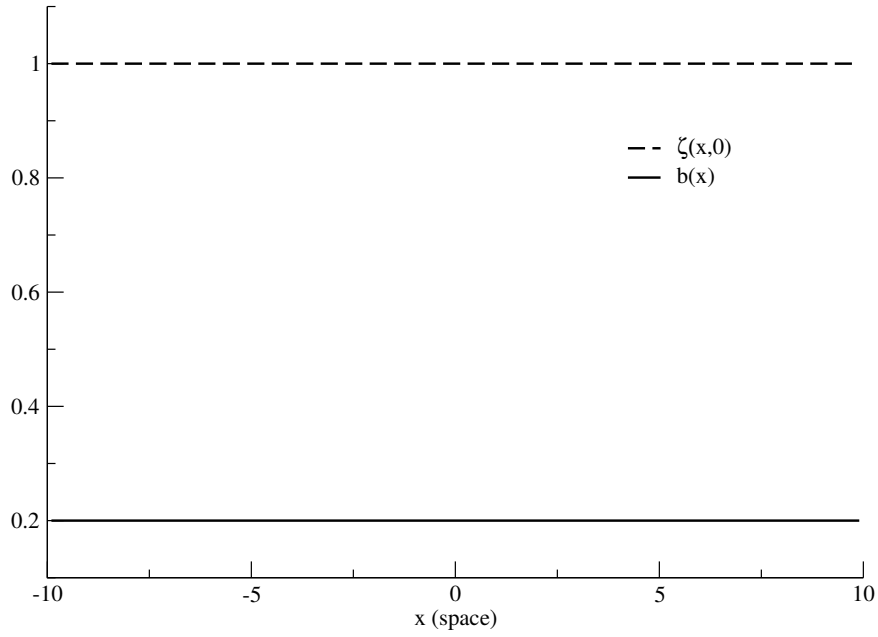
FIG. 5. Profile $\zeta(x,t)$ of the starting initial condition to initialize the global optimization procedure.

boundary conditions and initial condition $q(x,t) = 0$, $h(x,0) = 1$ and $b(x) = 0.2$, see Figure 5. We use the propose methodology for finding the set of condition to find the sought profile up to the final time $t_{out} = 3$. For simulations we have used 300 cells, $C_{cfl} = 0.1$, $\varepsilon = 0.01$. To update the sought bottom surface we set $\lambda_{IP} = 1.9$. Figure 6 shows the result of the simulation, as you can see the system achieve the sough profile with a good resolution despite of the few iterations as well. Notice that $\bar{\zeta}$ does not correspond to any exact solution of model (4..13) with initial conditions used in this model. So this is not a test using synthetic data. However, the procedure is able to provide an acceptable approximation of the sought profile. Model system consisting of (4..13), (4..14) and (4..15), is solved using the scheme of reference, appendix A. Notice also that this strategy is independent of the elected functional J and can be employed to deal with more general shallow water regimes [16, 23, 33], including the possibility of detecting moving bottoms $b(x,t)$.

5. Conclusions

In the present paper, we have expressed the adjoint method for PDE-constrained optimization problems in a unified frame. A numerical scheme based on a class of high order finite volume schemes has been adapted for solving the unified system. The scheme is able to deal with non-conservative equations and we have seen that the degree of approximation is at least those of conventional solvers. We have solved two test problems, the first one corresponds to the Burgers equations and the second one, the Nonlinear Shallow Water equations. In the case of Burgers equations, we assess two cases. First,
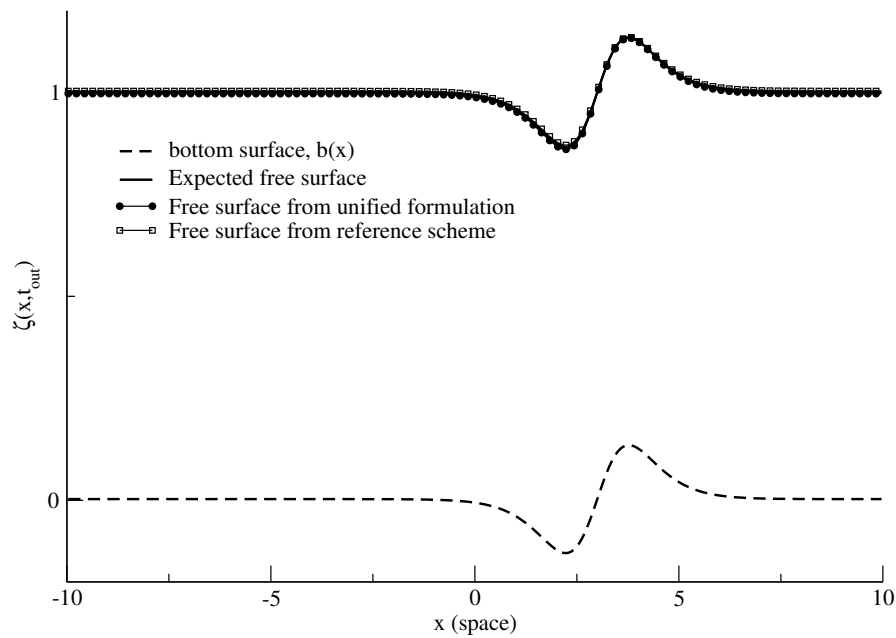
FIG. 6. Shallow water. Sought profile for the shallow water, $\varepsilon = 0.01$. Using 300 cells, $\lambda_{IP} = 1.9$. Bottom profile (thick line), free surface from unified formulation (circle), free surface from reference scheme (square) and expected free surface (continuous line). At 40*th* iteration, unified formulation generates $Error = 4.93 \cdot 10^{-3}$ and conventional formulation generate $Error = 5.49 \cdot 10^{-3}$.
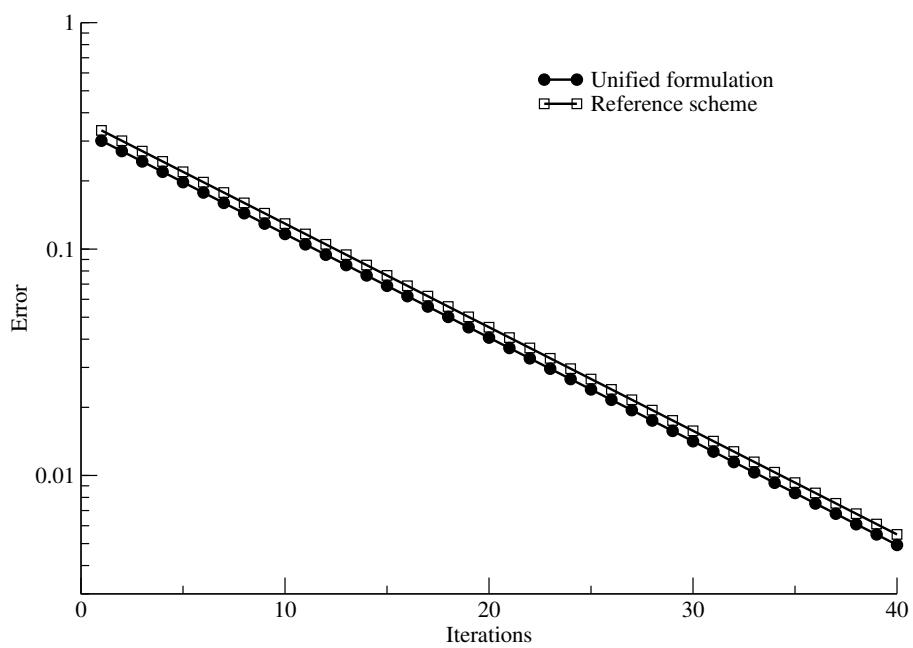
FIG. 7. Shallow water. Error measured as the distance with respect to the prescribed free surface for both, the unified formulation (square) and conventional formulation (circles).

we have solved the problem of finding an initial condition which is a discontinuous function, which generates a shock wave. The method finds the right location and also the main shape of the initial profile. We observed stationary overshot and undershot at the interface position, it is also observed by the reference scheme. Second, we recover a smooth initial condition, the constraint PDE is solved up to $T = 0.12$, before discontinuities start developing, so the solution remains continuous. In this case, the procedure has provided very good agreements in both the present and reference schemes. Regarding the shallow water equations, we also have obtained very good agreements. We have shown the evolution of the error between the computed sought parameters and that obtained in the simulations. We have observed that the present scheme always generated approximations with slightly better approximation than the reference scheme in the case of discontinuous solutions and for smooth cases, it has generated approximations whose accuracy is almost one order of magnitude.

Acknowledgements

A The numerical scheme of reference

In this section, we modify a strategy available in the literature for solving this type of problem, given by [26]. The original scheme is implemented to solve an inverse problem to find information at boundaries. In this work we are interested in solving initial value problems, boundary conditions in our case are just of the transmissive type. Moreover, we set the strategy for finding model parameters $\mathbf{b}$, through out the adjoint method. So, we provide an initial parameter $\mathbf{b}$ and then we apply the following procedure.

1) Given $\mathbf{b}^k$, we solve the hyperbolic model (2..1). Tests in this paper are conservatives, we neglect source terms $\mathbf{L}$. So, we use the following conservative scheme for marching in time up to the output time $t = T$, which is reached in a finite number of steps, let say n_T, that means starting from $t^0 = 0$, we do $t^{n+1} = t^n + \Delta t$, n_T times up to $t^{n_T} = T$.

$$
\begin{aligned}
\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \tfrac{\Delta t}{\Delta x}\left[\mathbf{R}_{i+\frac{1}{2}} - \mathbf{R}_{i-\frac{1}{2}}\right] + \Delta t\,\mathbf{L}(\mathbf{U}_i^n, \mathbf{b}_i^k) \\
+ \tfrac{\Delta t}{2\Delta x}\tilde{\mathbf{B}}(\mathbf{U}_i^n, \mathbf{b}_i^k)(\mathbf{b}_{i+1}^k - \mathbf{b}_{i-1}^k)\,,
\end{aligned}
\tag{A.1}
$$

where $\mathbf{R}_{i+\frac{1}{2}} = \frac{1}{2}(\mathbf{R}(\mathbf{U}_{i+1}^n, \mathbf{b}^k) + \mathbf{R}(\mathbf{U}_i^n, \mathbf{b}^k)) - \frac{\lambda_{i+\frac{1}{2}}}{2}(\mathbf{U}_{i+1}^n - \mathbf{U}_i^n)$. This corresponds to the well known numerical flux of Rusanov, $\lambda_{i+\frac{1}{2}} = \max(\lambda(\mathbf{U}_i^n), \lambda(\mathbf{U}_{i+1}^n))$ with $\lambda(\mathbf{U})$ being the maximum of the eigenvalues (in magnitude) of the Jacobian matrix of $\mathbf{R}$ with respect to $\mathbf{U}$.

2) Once the previous step is completed. We start the backward evolution of the so called adjoint method. We use a finite difference approach. We set the initial condition $\mathbf{P}_i^0 = 0$.

$$
\mathbf{P}_i^n = \mathbf{P}_i^{n+1} + \tfrac{\Delta t}{2\Delta x}\mathbf{J}_R^T(\mathbf{U}_i^{n+1})\left[\mathbf{P}_{i+1}^{n+1} - \mathbf{P}_{i-1}^{n+1}\right] - \Delta t\,\tilde{\mathbf{S}}(\mathbf{U}_i^{n+1})\,.
\tag{A.2}
$$

Index $n+1$ is consistent with the backward evolution in time, that means starting from $t^0 = T$, we do $t^{n+1} = t^n - \Delta t$. So in a finite number of steps, n_T, we reach $t^{n_T} = 0$.

3) Update the parameter $\mathbf{b}^k$. It is carried out using the gradient of the cost functional (2..2). The expression for the gradient of this functional depends on the problem at hand. In this paper they normally have the form $\nabla J_i = \mathbf{P}_i^{n_T}$ in terms of the discretization. Notice that this procedure can have more general structures and may also depend on the state variables, so let us express the gradient formally in terms of a functional, $G(\mathbf{U}, \mathbf{P}, \mathbf{b})$. So the update of the sought parameters has the form

$$\mathbf{b}_i^{k+1} = \mathbf{b}_i^k - \lambda_{IP} \cdot G(\mathbf{U}, \mathbf{P}, \mathbf{b}^k)_i \,, \tag{A.3}$$

where λ_{IP} is a prescribed constant value. Notice that, the procedure is carried out for finding parameters. However, for another type of problems, for example for finding initial conditions, the procedure is quite similar to the present one.

4) Stop the procedure using some stopped criterion. This is normally carried in terms of some relative error.

Let us point out that the scheme in [26] is globally implicit, in turns, here we derive the global explicit version of it. This is because the proposed scheme in this paper is globally explicit.

References

[1] F. Abergel and R. Temam. On some control problems in fluid mechanics. *Theoretical and Computational Fluid Dynamics*, 1(6):303–325, Nov 1990.

[2] Thomas Alazard, Pietro Baldi, and Daniel Han-Kwan. Control of water waves. *arXiv preprint arXiv:1501.06366*, 2015.

[3] Paul Andersson and Martin Berggren. *Computational Methods for Optimal Design and Control: Proceedings of the AFOSR Workshop on Optimal Design and Control Arlington, Virginia 30 September-3 October, 1997*. Progress in Systems and Control Theory 24, Jeff Borggaard, John Burns, Eugene Cliff, Scott Schreck (eds.). Birkhäuser Basel, 1 edition, 1998.

[4] Joseph Boussinesq. Théorie des ondes et des remous qui se propagent le long d'un canal rectangulaire horizontal, en communiquant au liquide contenu dans ce canal des vitesses sensiblement pareilles de la surface au fond. *Journal de Mathématiques Pures et Appliquées*, pages 55–108, 1872.

[5] C. Castro, F. Palacios, and E. Zuazua. An alternating descent method for the optimal control of the inviscid Burgers equation in the presence of shocks. *Mathematical Models and Methods in Applied Sciences*, 18(3):369–416, 2008.

[6] C. E. Castro and E. F. Toro. Solvers for the high–order Riemann problem for hyperbolic balance laws. *Journal of Computational Physics*, 227:2481–2513, 2008.

[7] Manuel J. Castro, José M. Gallardo, and Carlos Parés. High order finite volume schemes based on reconstruction of states for solving hyperbolic systems with nonconservative products. applications to shallow-water systems. *Math. Comput.*, 75(255):1103–1134, 2006.

[8] A. Barré de Saint-Venant. Théorie du mouvement non permanent des eaux, avec application aux crues des rivières et à lintroduction des marées dans leur lit. *Comptes Rendus des séances de l'Académie des Sciences*, 73:237–240, 1871.

[9] M. Dumbser and E. F. Toro. A simple extension of the Osher Riemann solver to non-conservative hyperbolic systems. *Journal of Scientific Computing*, 48:70–88, 2010.

[10] M. Dumbser, O. Zanotti, R. Loubere, and S. Diot. A posteriori subcell limiting of the discontinuous Galerkin finite element method for hyperbolic conservation laws. *Journal of Computational Physics*, 278:47–75, 2014.

[11] K. O. Friedrichs and P. D. Lax. Systems of conservation equations with a convex extension. *Proceedings of the National Academy of Sciences of the United States of America*, 68:1686–1688, 1971.

[12] S. K. Goduvov. An interesting class of quasilinear systems. *Doklady Akademii nauk SSSR*, 139:521–523, 1961.

[13] Max D. Gunzburger. *Perspectives in flow control and optimization*. Advances in design and control. Society for Industrial and Applied Mathematics, 2003.

[14] A. Harten and S. Osher. Uniformly High-Order Accurate Nonoscillatory Schemes. I. *SIAM Journal on Numerical Analysis*, 24(2):279–309, 1987.

[15] Michael Hinze, Rene Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE Constraints*. Mathematical Modelling: Theory and Applications. Springer, 1 edition, 2008.

[16] Samer Israwi. Large time existence for 1d green-naghdi equations. *Nonlinear Analysis: Theory, Methods & Applications*, 74(1):81–93, 2011.

[17] F. James and M. Postel. Numerical gradient methods for flux identification in a system of conservation laws. *Journal of Engineering Mathematics*, 60(3-4):293–317, 2008.

[18] F. James and M. Sepúlveda. Convergence results for the flux identification in a scalar conservation law. *SIAM Journal on Control and Optimization*, 37(3):869–891, 1999.

[19] RS Johnson. *A modern introduction to the mathematical theory of water waves*, volume 19. Cambridge University Press, 1997.

[20] H. Kang and K. Tanuma. Inverse problems for scalar conservation laws. *Inverse Problems*, 21:1047–1059, 2005.

[21] D.A. Knopoff, D.R. Fernndez, G.A. Torres, and C.V. Turner. Adjoint method for a tumor growth pde-constrained optimization problem. *Computers & Mathematics with Applications*, 66(6):1104 – 1119, 2013.

[22] D. Lannes. Well-posedness of the water-waves equations. *Journal of the American Mathematical Society*, 18(3):605–654, 2005.

[23] D. Lannes. *The Water Waves Problem: Mathematical Analysis and Asymptotics*. Mathematical Surveys and Monographs. American Mathematical Society, 2013.

[24] Peter D Lax. Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Communications on pure and applied mathematics*, 7(1):159–193, 1954.

[25] R. Lecaros and R. Zuazua. Control of 2d scalar conservation laws in the presence of shocks. *Mathematics of computation. Submitted*, 2014.

[26] J.-M. Lellouche, J.-L. Devenon, and I. Dekeyser. Boundary control of burgers' equationa numerical approach. *Computers & Mathematics with Applications*, 28(5):33 – 44, 1994.

[27] G. Montecinos, C. E. Castro, M. Dumbser, and E. F. Toro. Comparison of solvers for the generalized Riemann problem for hyperbolic systems with source terms. *Journal of Computational Physics*, 231:6472–6494, 2012.

[28] Jürgen Moser. A rapidly convergent iteration method and non-linear partial differential equations-i. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 20(2):265–315, 1966.

[29] Hayk Nersisyan, Denys Dutykh, and Enrique Zuazua. Generation of 2D water waves by moving bottom disturbances. *IMA Journal of Applied Mathematics*, 80(4):1235–1253, 2015.

[30] Hayk Nersisyan, Denys Dutykh, and Enrique Zuazua. Generation of 2d water waves by moving bottom disturbances. *IMA Journal of Applied Mathematics*, 80(4):1235–1253, 2015.

[31] S. Osher and F. Solomon. Upwind difference schemes for hyperbolic systems of conservation laws. *Mathematics of Computation*, 38(158):339–374, apr 1982.

[32] P. Parés. Numerical methods for nonconservative hyperbolic systems: a theoretical framework. *SIAM Journal on Numerical Analysis*, 44(1):300–321, 2006.

[33] D. H. Peregrine. Long waves on a beach. *Journal of Fluid Mechanics*, 27(2):815–827, 1967.

[34] Denis Serre and Alexis F Vasseur. About the relative entropy method for hyperbolic systems of conservation laws. 2015.

[35] M. Taylor. *Partial Differential Equations III: Nonlinear Equations*. Applied Mathematical Sciences. Springer New York, 2010.

[36] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction*. Springer-Verlag, third edition, 2009. ISBN 978-3-540-25202-3.

[37] E. F. Toro, R. C. Millington, and L. A. M. Nejad. Towards very high–order Godunov schemes. In *Godunov Methods: Theory and Applications. Edited Review, E. F. Toro (Editor)*, pages 905–937. Kluwer Academic/Plenum Publishers, 2001.

[38] E. F. Toro and V. A. Titarev. Solution of the generalised Riemann problem for advection–reaction equations. *Proceedings of the Royal Society of London A*, 458:271–281, 2002.

[39] S. Ulbrich. A sensitivity and adjoint calculus for discontinuous solutions of hyperbolic conservation laws with source terms. *SIAM Journal on Control and Optimization*, 41(3):740–797, March 2002.