HAL open science

Numerical analysis of quadratized schemes. Application to the simulation of the nonlinear piano string

Guillaume Castera, Juliette Chabassier

▶ To cite this version:

Guillaume Castera, Juliette Chabassier. Numerical analysis of quadratized schemes. Application to the simulation of the nonlinear piano string. RR-9516, Inria. 2023. hal-04182465

HAL Id: hal-04182465

https://hal.science/hal-04182465v1

Submitted on 19 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Numerical analysis of quadratized schemes. Application to the simulation of the nonlinear piano string

Guillaume Castera, Juliette Chabassier

RESEARCH REPORT
N° 9516
September 2023
Project-Team MAKUTU

ISSN 0249-6399 ISRN INRIA/RR--9516--FR+ENG

Numerical analysis of quadratized schemes. Application to the simulation of the nonlinear piano string

Guillaume Castera*, Juliette Chabassier*

Project-Team MAKUTU

Research Report n° 9516 — September 2023 — 54 pages

Abstract: In this report, we present energy quadratization techniques for a Hamiltonian system of nonlinear wave equations formulated at order 2 in time. On a generic system, we present the so-called Invariant Energy Quadratization (IEQ) and Scalar Auxiliary Variable (SAV) methods as well as their energy conservation properties and discretization strategies in space and time. Unlike the iterative techniques commonly used for nonlinear systems to guarantee certain invariances, these two methods lead to algorithms whose complexity is known in advance and rely on the simple inversion of a linear system at each time step. In spite of an unconditional stability and an attractive complexity, the literature mentions problematic application cases with an uncontrolled accuracy.

The numerical properties (stability, consistency and uniform convergence in time with respect to the CFL) of schemes obtained by hybridization between θ-scheme and quadratization are studied for two classes of nonlinear terms: a nonlinearity concerning the solution field, and a nonlinearity concerning its gradient.

These results are then applied to a geometrically exact nonlinear piano string for which numerical results are presented. The influence of the discretization parameters is studied and related to the theoretical results. The choices for the best accuracy or computational cost are illustrated. Some parameters can induce the space-time non-convergence of the schemes for a nonlinearity on the gradient, as it is the case for the piano string.

Key-words: Numerical analysis, Space-Time convergence, Invariant Energy Quadratization, Scalar Auxiliary Variable

* Makutu team, Inria Bordeaux - Sud-Ouest

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 avenue de la Vieille Tour
33405 Talence Cedex

Analyse numérique des schémas quadratisés. Application à la simulation de la corde de piano non-linéaire

Résumé : Dans ce rapport, nous présentons en détail des techniques de quadratisation d'énergie pour un système hamiltonien d'équations d'ondes non-linéaires formulé à l'ordre 2 en temps. Sur un système générique, nous présentons les méthodes dites d'Invariant Energy Quadratization (IEQ) et de Scalar Auxiliary Variable (SAV) ainsi que leurs propriétés de conservation d'énergie et des stratégies de discrétisation en espace et en temps. Contrairement aux techniques itératives couramment utilisées pour des systèmes non-linéaires afin de garantir certaines invariances, ces deux méthodes conduisent à des algorithmes de résolution dont la complexité est connue à l'avance, et reposent sur la simple inversion d'un système linéaire à chaque pas de temps. Malgré une stabilité inconditionnelle et une complexité attrayante, la bibliographie mentionne des cas d'application problématiques à la précision non maitrisée.
Les propriétés numériques (stabilité, consistance et convergence en temps uniformément par rapport à la CFL) des schémas obtenus par hybridation entre θ-schéma et quadratisation sont étudiées pour deux classes de non-linéarités : une non-linéarité portant sur le champs solution, et une non-linéarité portant sur son gradient.
Ces résultats sont alors appliqués à une corde de piano non-linéaire géométriquement exacte pour laquelle sont présentés des résultats de simulations numériques. L'influence des paramètres de discrétisation est étudiée et mise en relation avec les résultats théoriques. Les choix garantissant la meilleure précision ou coût de calcul sont illustrés. Certains paramètres peuvent notamment induire la non-convergence espace-temps des schémas pour une non-linéarité portant sur le gradient, comme c'est le cas pour la corde de piano.

Mots-clés : Analyse numérique, Convergence spatio-temporelle, Quadratisation à Energie Invariante, Variable Auxiliaire Scalaire

Contents

Introduction **5**

I Abstract problem **7**

1 General equations **7**

2 Energy Quadratization **8**
2.1 Invariant Energy Quadratization (IEQ) . . . 8
2.1.1 Variational formulation and Energy . . . 9
2.1.2 Spatial discretization . . . 10
2.1.3 Time discretization . . . 11
2.1.4 Matrix formulation . . . 12
2.1.5 Practical solution of the scheme . . . 12
2.2 Scalar Auxiliary Variable (SAV) . . . 14
2.2.1 Variational formulation and Energy . . . 14
2.2.2 Spatial discretization . . . 15
2.2.3 Time discretization . . . 15
2.2.4 Matrix formulation . . . 16
2.2.5 Practical solution of the scheme . . . 16

3 Numerical analysis **18**
3.1 Stability . . . 18
3.2 Discrete regularity assumptions . . . 21
3.3 Consistency . . . 25
3.4 Convergence for type 1 nonlinear terms . . . 27
3.5 Convergence for type 2 nonlinear terms . . . 29

4 Phase formulation of quadratized schemes **30**
4.1 Phase P-IEQ numerical scheme . . . 30
4.2 Phase P-SAV numerical scheme . . . 31

II Simulation of the nonlinear piano string **32**

1 Piano string model **32**

2 Properties of the nonlinear function **33**

3 Numerical schemes for the piano string **34**
3.1 Invariant Energy Quadratization (IEQ) . . . 34
3.2 Scalar Auxiliary Variable (SAV) . . . 37

4 Numerical results **38**
4.1 Solutions of the schemes . . . 39
4.2 Energy preservation . . . 40
4.3 Time convergence . . . 40
4.4 Aliasing problems . . . 42
4.5 Space-Time convergence . . . 42
4.5.1 Unconditionally stable scheme . . . 42
4.5.2 Conditionally stable scheme . . . 43
4.6 Long time simulations . . . 44
4.7 Influence of the α-decomposition . . . 46

4.7.1 Convergence constant estimation 46
4.7.2 Lipschitz constant estimation 47
4.7.3 Influence of the θ-scheme 48
4.8 Influence of the auxiliary constant c 49
4.9 Choice of the parameters 50

5 Conclusions and Prospects **51**

Appendices **52**

A Discrete Gronwall Lemma **52**

Introduction

Nonlinear wave equations are quite frequently encountered in many application domains like acoustics, fluid and solid state, optics, quantum, and others. We consider in this work Hamiltonian wave equations, which can entirely be described from the knowledge of a Hamiltonian function. These systems have energy conservation properties that can be exploited to perform the mathematical analysis of the equations.
Such a property is also very useful for numerical computation as it ensures the preservation of a discrete analog of the energy identity which allows to derive a-posteriori stability estimates and convergence results of the time integration scheme for a lot of application cases [Joly, 2003], [Chabassier and Imperiale, 2017]. These estimates can especially be used to couple multiple systems even if each one of them has a different integration strategy.

The system of equations that we consider here consists in finding $u \equiv u(x,t) \in \mathbb{R}^p$ for $x \in \Omega$, $t \in [0,T]$, such that

$$\partial_t^2 M u - \mathrm{Div}\,(A\nabla u + \nabla F_2(\nabla u)) + \nabla F_1(u) = f \tag{.1}$$

with $\Omega \subset \mathbb{R}^d$, $F_1 : \mathbb{R}^p \to \mathbb{R}$ and $F_2 : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ two nonlinear forms, and M, A two linear operators of $\mathbb{R}^p \times \mathbb{R}^p$.
$u : \Omega \times [0,T] \to \mathbb{R}^p$ is a function of space and time and $f \in L^2(\Omega)$ an exterior source term.

Two separated groups stand out among all the nonlinear terms that can be found in physics. In this work we will call "type 1 nonlinear terms" the nonlinear terms which are a function of the unknown field u, for example Sine-Gordon equation [Rubinstein, 1970], [Barone et al., 1971] or Allen–Cahn and Cahn-Hilliard equation [Allen and Cahn, 1979], [Cahn and Hilliard, 1958], [Shen and Yang, 2010]. Piano hammer impacting a string [Chabassier, 2012], [Bilbao et al., 2015] or wind instrument's reeds [Bilbao et al., 2015] can also be modeled with type 1 nonlinear functions. We will call "type 2 nonlinear terms" the nonlinear terms which are a function of the gradient of the unknown field ∇u, for example a string, plate, or 3D solid element submitted to large deformations [Banks et al., 1995]. In the context of wave equations, a function like F_1 is of type 1 and the F_2 function is of type 2.

The existence and regularity of continuous solutions to (.1) is not the topic of this work. Such results will instead be supposed in order to perform the stability and convergence analysis of some numerical integration strategies.

A widely spread strategy to solve nonlinear Hamiltonian equations with energy consistent methods is the use of Discrete Gradients. In [He and Sun, 2020], [Rincon and Quintino, 2016] such schemes are used in dimensions 2 and 3 for type 1 nonlinear term. [Shen and Yang, 2010] uses it for Allen-Cahn and Cahn-Hilliard equations and give some convergence proofs. [Gonzalez, 2000] deals with nonlinear elasticity and in [Bilbao et al., 2015], [Chatziioannou and Van Walstijn, 2015], a Discrete Gradient approach is used to tackle contact terms in musical acoustics, which can be of both types 1 and 2. In [Chabassier and Joly, 2010] such schemes are used for the 1D nonlinear piano string which is of type 2.

The discrete gradient schemes have the major drawback of leading to implicit nonlinear schemes that must be solved with iterative techniques. It requires to choose a convergence threshold, and induces a consequent number of iterations depending on how "hard" the nonlinear term is. The computation cost is often very high and unpredictable, as well as the implementation effort.

Recent strategies have appeared that guarantee a discrete energy identity while increasing efficiency. The so-called Invariant Energy Quadratization (IEQ) schemes were introduced in [Yang, 2016], [Zhao et al., 2017] in the context of phase-fields models (nonlinearities of type 1). The so-called Scalar Auxiliary Variable (SAV) schemes introduced in [Shen et al., 2019] and all its variants [Liu and Li, 2022], [Liu and Li, 2020] were applied for gradient flows, but also for incompressible Navier-Stokes [Lin et al., 2019], Sine-Gordon [Jiang et al., 2019], or general Hamiltonian equations [Cai and Shen, 2020, Jiang et al., 2021, Li and Sun, 2020], all of them with nonlinearities of type 1. These techniques were applied recently to the geometrically exact piano string which is of type 2 in [Ducceschi and Bilbao, 2019], [Ducceschi and Bilbao, 2022], [Ducceschi et al., 2022] on the discrete ordinary differential equation (ODE) system obtained after space discretization with finite differences.

Although the literature for applied cases is very rich and creative, numerical analysis of these schemes is quite scarce. A space-time convergence proof can be found in [Jiang et al., 2019] for a type 1 Sine-Gordon nonlinear wave equation discretized in space with finite differences, and [Yang and Zhang, 2020], [Shen and Xu, 2018] study both IEQ and SAV convergence in the context of diffusive phase-fields under the form of gradient flow models, with nonlinear terms that seem to behave theoretically like type 1 terms. To the best of our knowledge, no mathematical study of these schemes with type 2 nonlinearities has been proposed yet.

It is a global agreement among all these previous references that these quadratized schemes are very efficient and easy to implement compared to other nonlinear schemes like convex splitting [Eyre, 1998] or discrete gradient. However, some authors point out some aliasing problems [Ducceschi and Bilbao, 2019] when using SAV with too large time steps even though it is unconditionally stable. [Shen et al., 2019], [Bilbao et al., 2023] also mention a stabilization parameter that can help (or not) the scheme to give precise results with the largest possible time steps, but no clear stabilization process is given.

This report is structured in two parts. In the first part we present linearly implicit and energy-preserving space-time discretizations of I.1 using spectral finite elements along with IEQ and SAV methods. The quadratization techniques and the numerical schemes are presented along with continuous and discrete energy balances and solving techniques.
Numerical analysis including stability, consistency and time-convergence proofs of the IEQ scheme for both nonlinear types 1 and 2 is given in section 3. It can easily adapt to SAV. Treating the case of nonlinear terms of type 2 requires a finite dimension assumption (fixed space step) which is not needed for type 1 nonlinear terms. The second part presents numerical results of these schemes for the geometrically exact piano string including efficiency comparisons, times and space-time convergence rates, and investigations on the stabilization methods mentioned above and on the influence of some other discretization parameters.

Notations

For U and V two vectors of $\mathbb{R}^n$ we denote '$\cdot$' as the component-wise inner product:

$$U \cdot V = \sum_{i=1}^{n} U_i V_i \tag{.2}$$

The associated norm is $\|.\|_2$

For A and B two matrices of $\mathcal{M}_n(\mathbb{R})$ we denote ':' as the matricial double dot product also known as Frobenuis dot product:

$$A : B = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j} B_{i,j} \tag{.3}$$

and

$$\|A\|_F = \sqrt{A : A} = \sqrt{Tr\left({}^t AA\right)} \tag{.4}$$

Part I
Abstract problem

1 General equations

Let $\Omega \subset \mathbb{R}^d$ measurable.
Let $F_1 : \mathbb{R}^p \to \mathbb{R}$ and $F_2 : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ be two nonlinear forms and M, A two linear positive operators of $\mathbb{R}^p \times \mathbb{R}^p$.
We consider $u : \Omega \times [0,T] \to \mathbb{R}^p$ a function of space and time and $f \in L^2(\Omega)$ such that

$$\partial_t^2 Mu - \mathrm{Div}\,(A\nabla u + \nabla F_2(\nabla u)) + \nabla F_1(u) = f \tag{I.1}$$

with Dirichlet boundary conditions on $\partial\Omega$ and initial conditions described by an initail function u_0:

$$\begin{cases} \forall (x,t) \in \partial\Omega \times [0,T], \quad u(x,t) = 0 & \text{(I.2a)} \\ \forall x \in \Omega, \quad u(x,0) = u_0 & \text{(I.2b)} \end{cases}$$

Assumption 1.1 (Existence and regularity of strong solution.)
Equation (I.1) *has a unique solution*

$$u \in \mathscr{C}^4\left([0,T], \left(L^2(\Omega)\right)^p\right) \cap \mathscr{C}^3\left([0,T], \left(H^1(\Omega)\right)^p\right)$$

Remark 1.1 *This result can be obtained for example with Hille-Yoshida theorem if the source term f is regular enough.*

Theorem 1.1 (Weak formulation)
We seek $u \in H^1(\Omega)^p$ so that

$$\forall u^* \in H^1(\Omega)^p, \quad \int_\Omega \partial_t^2 Mu \cdot u^* + \int_\Omega A\nabla u : \nabla u^* + \int_\Omega \nabla F_2(\nabla u) : \nabla u^* + \int_\Omega \nabla F_1(u) \cdot u^* = \int_\Omega f \cdot u^* \tag{I.3}$$

Theorem 1.2 (Energy Conservation Identity)
Any solution to I.3 satisfies

$$\frac{d\mathcal{E}_0}{dt} = \int_\Omega f \cdot \partial_t u \tag{I.4}$$

with

$$\mathcal{E}_0(t) = \frac{1}{2}\int_\Omega M\partial_t u \cdot \partial_t u + \frac{1}{2}\int_\Omega A\nabla u : \nabla u + \int_\Omega F_2(\nabla u) + \int_\Omega F_1(u) \tag{I.5}$$

Proof *Just apply the weak formulation* (I.3) *with $u^* = \partial_t u \in H^1(\Omega)^p$ from assumption 1.1.*
Note that this energy is not a quadratic form, and in order to be positive F_1 and F_2 must be positive.

Remark 1.2 *Notice that the choice of F_2 is not unique. We can introduce a matrix α such that $A = \alpha A + (I_p - \alpha)A$ and transform the equation into*

$$\partial_t^2 Mu - Div(\alpha A\nabla u + \nabla F_{2,\alpha}(\nabla u)) + \nabla F_1(u) = f \tag{I.6}$$

with $F_{2,\alpha}(\nabla u) = \frac{1}{2}(I_p - \alpha)A\nabla u : \nabla u + F_2(\nabla u)$.

This α*-decomposition is called stabilization in [Shen et al., 2019], [Bilbao et al., 2023] and allows to adjust the amount of linear terms to put inside* F_2*. With* $\alpha = I_p$ *no part of the linear term is in* $F_{2,\alpha}$*, but the closer* α *is from 0 the more linear terms go into* $F_{2,\alpha}$*.*
In part II about the simulation of the piano string we will explore the tuning of α *and see that its value has a huge influence on the accuracy of the quadratized schemes.*

2 Energy Quadratization

If F_1 and F_2 are bounded below, we can find two real constants c_1 and c_2 so that $c_i + 2F_i$ is positive. And then we can do the following transformation:

$$\frac{d\mathcal{E}}{dt} = \frac{d\mathcal{E}_0}{dt} = \int_\Omega f \cdot \partial_t u \tag{I.7}$$

with

$$\mathcal{E}(t) = \mathcal{E}_0(t) + \frac{c_1 + c_2}{2} = \frac{1}{2}\int_\Omega M\partial_t u \cdot \partial_t u + \frac{1}{2}\int_\Omega A\nabla u : \nabla u + \frac{1}{2}\left(c_2 + 2\int_\Omega F_2(\nabla u)\right) + \frac{1}{2}\left(c_1 + 2\int_\Omega F_1(u)\right) \tag{I.8}$$

Now since $c_i + 2F_i$ is positive we can introduce an auxiliary variable which is a square root of it and it will lead to the so-called Invariant Energy Quadratization method (IEQ). Taking a square root of the entire $c_i + 2\int_\Omega F_i$ as an auxiliary variable is also possible and will lead to the Scalar Auxiliary Variable method (SAV).
The preserved quantity of the resulting system of equations is no longer the energy but a modified quadratic invariant.

Assumption 2.1 (Local Lipschitz assumptions on functions F_i)
We asses that there exists two subsets $\mathcal{I}$ *and* $\mathcal{I}_\nabla$ *such that*

$$\begin{cases} \forall (a,b) \in \mathcal{I}^2, \quad |F_1(a) - F_1(b)| \leqslant C_{f1} \|a-b\|_2 & \text{(I.9a)} \\ \forall (a,b) \in \mathcal{I}^2, \quad \|\nabla F_1(a) - \nabla F_1(b)\|_2 \leqslant C_{df1} \|a-b\|_2 & \text{(I.9b)} \\ \forall (a,b) \in \mathcal{I}_\nabla^2, \quad |F_2(a) - F_2(b)| \leqslant C_{f2} \|a-b\|_F & \text{(I.9c)} \\ \forall (a,b) \in \mathcal{I}_\nabla^2, \quad \|\nabla F_2(a) - \nabla F_2(b)\|_F \leqslant C_{df2} \|a-b\|_F & \text{(I.9d)} \end{cases}$$

2.1 Invariant Energy Quadratization (IEQ)

Assumption 2.2 *Let* u *be the solution of I.3.*
We assume that there exist $(\beta_1, \beta_2) \in (\mathbb{R}_+^*)^2$ *depending on the source* f *such that for all* $(x,t) \in \Omega \times [0,T]$
$-\frac{\beta_1}{2} < F_1(u(x,t))$ *and* $-\frac{\beta_2}{2} < F_2(\nabla u(x,t))$.

It ensures that $2F_1(u(x,t)) + c_1 > c_1 - \beta_1 > 0$ *and* $2F_2(\nabla u(x,t)) + c_2 > c_2 - \beta_2 > 0$

Let's introduce two auxiliary variables z_1 and z_2:

$$\forall (x,t) \in \mathbb{R}^p \times [0,T], \quad \begin{cases} z_1(x,t) \equiv \sqrt{2F_1(u(x,t)) + c_1} & \text{(I.10a)} \\ z_2(x,t) \equiv \sqrt{2F_2(\nabla u(x,t)) + c_2} & \text{(I.10b)} \end{cases}$$

with two constants c_1 and c_2 so that the radicands are strictly positive, and also two auxiliary functions g_1 and g_2 the functions such that:

$$\begin{cases} \forall q \in \mathbb{R}^p, \quad g_1(q) = \dfrac{1}{\sqrt{2F_1(q)+c_1}} \nabla F_1(q) \quad \in \mathbb{R}^d & \text{(I.11a)} \\ \forall q \in \mathbb{R}^p \times \mathbb{R}^d, \quad g_2(q) = \dfrac{1}{\sqrt{2F_2(q)+c_2}} \nabla F_2(q) \quad \in \mathbb{R}^p \times \mathbb{R}^d & \text{(I.11b)} \end{cases}$$

and the operators G_1 and G_2:

$$\begin{cases} \forall u \in H^1(\Omega)^p, \quad G_1(u) = \dfrac{1}{\sqrt{2F_1(u)+c_1}} \nabla F_1(u) & \text{(I.12a)} \\ \forall u \in H^1(\Omega)^p, \quad G_2(\nabla u) = \dfrac{1}{\sqrt{2F_2(\nabla u)+c_2}} \nabla F_2(\nabla u) & \text{(I.12b)} \end{cases}$$

so that for $(x,t) \in \Omega \times [0,T]$, $G_i(u)(x,t) = g_i(u(x,t))$.

We can rewrite (I.1) as:
Seek $u : \Omega \times [0,T] \to \mathbb{R}^d$ and $z_i : \Omega \times [0,T] \to \mathbb{R}$ such that

$$\begin{cases} \partial_t^2 M u - \mathrm{Div}\,(A\nabla u + z_2 G_2(\nabla u)) + z_1 G_1(u) = f & \text{(I.13a)} \\ \partial_t z_1 = G_1(u) \cdot \partial_t u & \text{(I.13b)} \\ \partial_t z_2 = G_2(\nabla u) : \partial_t \nabla u & \text{(I.13c)} \end{cases}$$

with the same conditions I.2 for u and an extra initial condition:

$$\forall x \in \Omega, \quad \begin{cases} z_1(x,0) = \sqrt{2F_1(u_0(x)) + c_1} & \text{(I.14a)} \\ z_2(x,0) = \sqrt{2F_2(\nabla u_0(x)) + c_2} & \text{(I.14b)} \end{cases}$$

Proposition 2.1
The two formulations (I.1) *and* (I.13) *with conditions I.2 and I.14 are equivalent in the sens of distributions.*

Proof *Let u be the solution of* (I.1) *with conditions I.2. Then u is also solution of* (I.13) *with auxiliary variables defined as I.10 and conditions I.2 and I.14.*
Now let (u, z_1, z_2) be the solution of (I.13) *with conditions I.2 and I.14. Equations* (I.13b) *and* (I.13c) *ensure that $\partial_t \left(z_i - \sqrt{2F_i + c_i}\right) = 0$ and because of conditions I.14 after time integration we have $z_i - \sqrt{2F_i + c_i} = 0$. Using this in I.13a along with the definitions of G_i functions gives back equation I.1.*

2.1.1 Variational formulation and Energy

Since u is in $H^1(\Omega)^p$ the variational spaces of z_1 and z_2 are restricted. We will seek z_1 in $H^1(\Omega)$ but because it is a function of the gradient of u, z_2 will only be in $L^2(\Omega)$.

Proposition 2.2 (Weak formulation of IEQ)
We seek $u \in H^1(\Omega)^p$, $z_1 \in H^1(\Omega)$ and $z_2 \in L^2(\Omega)$ so that for all $u^* \in H^1(\Omega)^p$ and all $z_1^* \in H^1(\Omega)$, $z_2^* \in L^2(\Omega)$:

$$\begin{cases} \displaystyle\int_\Omega M\partial_t^2 u \cdot u^* + \int_\Omega A\nabla u : \nabla u^* + \int_\Omega z_2 G_2(\nabla u) : \nabla u^* + \int_\Omega z_1 G_1(u) \cdot u^* = \int_\Omega f \cdot u^* & \text{(I.15a)} \\ \displaystyle\int_\Omega \partial_t z_1 z_1^* = \int_\Omega z_1^* G_1(u) \cdot \partial_t u & \text{(I.15b)} \\ \displaystyle\int_\Omega \partial_t z_2 z_2^* = \int_\Omega z_2^* G_2(\nabla u) : \partial_t \nabla u & \text{(I.15c)} \end{cases}$$

We can also write this with weak formulation with functional forms:

$$\begin{cases} m(\partial_t^2 u, u^*) + a(u, u^*) + \bar{G}_1(u, u^*, z_1) + \bar{G}_2(u, u^*, z_2) = f(u^*) & \text{(I.16a)} \\ \ell(\partial_t z_1, z_1^*) = \bar{G}_1(u, \partial_t u, z_1^*) & \text{(I.16b)} \\ \ell(\partial_t z_2, z_2^*) = \bar{G}_2(u, \partial_t u, z_2^*) & \text{(I.16c)} \end{cases}$$

where for u and v in $H^1(\Omega)^p$

$$\begin{cases} m(u, v) = \oint_\Omega Mu \cdot v & \text{(I.17a)} \\ a(u, v) = \oint_\Omega A\nabla u : \nabla v & \text{(I.17b)} \\ f(v) = \oint_\Omega f \cdot v & \text{(I.17c)} \end{cases}$$

and for z and z^* in $L^2(\Omega)$

$$\ell(z, z^*) = \oint_\Omega zz^* \tag{I.18}$$

For u and v in $H^1(\Omega)^p$ and z in $L^2(\Omega)$ we also denoted

$$\begin{cases} \bar{G}_1(u, v, z) = \oint_\Omega zG(u) \cdot v & \text{(I.19a)} \\ \bar{G}_2(u, v, z) = \oint_\Omega zG(\nabla u) : \nabla v & \text{(I.19b)} \end{cases}$$

G_i is a nonlinear form in its first variable u, and linear in v and z.

Theorem 2.2 (Energy conservation of IEQ formulation)
Any solution to I.16 satisfies

$$\frac{d\mathcal{E}}{dt} = f(\partial_t u) = \int_\Omega f \cdot \partial_t u \tag{I.20}$$

with

$$\mathcal{E}(t) = \frac{1}{2}m(\partial_t u, \partial_t u) + \frac{1}{2}a(u, u) + \frac{1}{2}\ell(z_1, z_1) + \frac{1}{2}\ell(z_2, z_2) = \mathcal{E}_0(t) + \frac{c_1 + c_2}{2} \tag{I.21}$$

Proof *Just apply the weak formulation* (I.16) *with* $u^* = \partial_t u \in H^1(\Omega)^p$ *and* $z_i^* = z_i \in L^2(\Omega)$.
Note that the energy is now a quadratic and positive form.

2.1.2 Spatial discretization

Let $(\mathcal{Q}_h)_{h>0}$, $(\mathcal{Z}_{h,1})_{h>0}$ and $(\mathcal{Z}_{h,2})_{h>0}$ be Galerkin conform approximations of $H^1(\Omega)^p$, $H^1(\Omega)$ and $L^2(\Omega)$ respectively.
We discretize them with High-Order Spectral Lagrange Finite Elements $\mathbb{P}_r$ on a conform mesh $\mathcal{T}_h$ of Ω with order $r \in \mathbb{N}^*$.
We also introduce $n_h^u = \dim \mathcal{Q}_h$, $n_h^{z1} = \dim \mathcal{Z}_{h,1}$ and $n_h^{z2} = \dim \mathcal{Z}_{h,2}$, and $(\varphi_i)_{1 \leqslant i \leqslant n_h^u}$ are the basis functions of $\mathcal{Q}_h$, and $(\phi_i)_{1 \leqslant i \leqslant n_h^{z1}}$ and $(\psi_i)_{1 \leqslant i \leqslant n_h^{z2}}$ the ones of $\mathcal{Z}_{h,1}$ and $\mathcal{Z}_{h,2}$ so that

$$u_h = \sum_{i=1}^{n_h^u} U_{h,i}\varphi_i, \qquad z_{h,1} = \sum_{i=1}^{n_h^{z1}} Z_{h,1,i}\phi_i, \qquad z_{h,2} = \sum_{i=1}^{n_h^{z2}} Z_{h,2,i}\psi_i \qquad \text{and} \qquad f_h = \sum_{i=1}^{n_h^u} F_{h,i}\varphi_i \tag{I.22}$$

with $U_h, Z_{h,1}, Z_{h,2}$ the vectorial representations in the finite element basis.

Assumption 2.3 (Convergence of the finite elements method.)
We suppose in the next sections that the constructed Galerkin approximation and finite elements are conform and that the discrete solution $(u_h, z_{h,1}, z_{h,2})$ converges to (u, z_1, z_2) in the sens that there exists $\delta_{h,u}$, $\delta_{h,z1}$ and $\delta_{h,z2}$ such that

$$\begin{cases} \|u - u_h\|_{H^1(\Omega)^p} \leqslant C\delta_{h,u} \underset{h\to 0}{\longrightarrow} 0 & \text{(I.23a)} \\ \|z_1 - z_{h,1}\|_{H^1(\Omega)} \leqslant C\delta_{h,z1} \underset{h\to 0}{\longrightarrow} 0 & \text{(I.23b)} \\ \|z_2 - z_{h,2}\|_{L^2(\Omega)} \leqslant C\delta_{h,z2} \underset{h\to 0}{\longrightarrow} 0 & \text{(I.23c)} \end{cases}$$

We can derive the semi-discretized variational formulation.
Find $u_h \in \mathcal{Q}_h$ and $z_{h,i} \in \mathcal{Z}_{h,i}$ such that for all $u_h^* \in \mathcal{Q}_h$ and all $z_{h,i}^* \in \mathcal{Z}_{h,i}$:

$$\begin{cases} m(\partial_t^2 u_h, u_h^*) + a(u_h, u_h^*) + \bar{G}_1(u_h, u_h^*, z_{h,1}) + \bar{G}_2(u_h, u_h^*, z_{h,2}) = f(u_h^*) & \text{(I.24a)} \\ \ell(\partial_t z_{h,1}, z_{h,1}^*) = \bar{G}_1(u_h, \partial_t u_h, z_{h,1}^*) & \text{(I.24b)} \\ \ell(\partial_t z_{h,2}, z_{h,2}^*) = \bar{G}_2(u_h, \partial_t u_h, z_{h,2}^*) & \text{(I.24c)} \end{cases}$$

2.1.3 Time discretization

For time discretization we use a constant time-step Δt so that $N\Delta t = T$.

Let's introduce some discrete operators δ, $\delta^{1/2}$, μ and $\mu^{1/2}$ such that:

$$\delta u_h = \frac{u_h^{n+1} - u_h^n}{\Delta t}, \quad \mu u_h = \frac{u_h^{n+1} + u_h^n}{2}, \quad \delta\mu\, u_h = \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} \quad \text{and} \quad \delta^2 u_h = \frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} \tag{I.25}$$

and

$$\delta^{1/2} z_h = \frac{z_h^{n+1/2} - z_h^{n-1/2}}{\Delta t} \quad \text{and} \quad \mu^{1/2} z_h = \frac{z_h^{n+1/2} + z_h^{n-1/2}}{2} \tag{I.26}$$

We also name $\{u_h\}_\theta^n = \theta u_h^{n+1} + (1-2\theta)u_h^n + \theta u_h^{n-1}$ the θ-scheme.

Like [Ducceschi and Bilbao, 2022] and [Ducceschi and Bilbao, 2019] we use an interleaved time grid to discretize the auxiliary variable.

Numerical Scheme 2.1 (IEQ Time Scheme)
We seek $(u_h, z_{h,1}, z_{h,2}) \in \mathcal{Q}_h \times \mathcal{Z}_{h,1} \times \mathcal{Z}_{h,2}$ so that for all $(u_h^*, z_{h,1}^*, z_{h,2}^*) \in \mathcal{Q}_h \times \mathcal{Z}_{h,1} \times \mathcal{Z}_{h,2}$ and all $n \in [\![0, N]\!]$

$$\begin{cases} m\left(\delta^2 u_h, u_h^*\right) + a\left(\{u_h\}_\theta^n, u_h^*\right) + \bar{G}_1\left(u_h^n, u_h^*, \mu^{1/2} z_{h,1}\right) + \bar{G}_2\left(u_h^n, u_h^*, \mu^{1/2} z_{h,2}\right) = f^n(u_h^*) & \text{(I.27a)} \\ \ell\left(\delta^{1/2} z_{h,1}, z_{h,1}^*\right) = \bar{G}_1\left(u_h^n, \delta\mu\, u_h, z_{h,1}^*\right) & \text{(I.27b)} \\ \ell\left(\delta^{1/2} z_{h,2}, z_{h,2}^*\right) = \bar{G}_2\left(u_h^n, \delta\mu\, u_h, z_{h,2}^*\right) & \text{(I.27c)} \end{cases}$$

Theorem 2.4 (Discrete energy Identity of IEQ scheme)
Any solution to I.27 satisfies

$$\delta^{1/2}\mathcal{E}_h = f^n\left(\delta\mu\, u_h\right) = \int_\Omega f_h^n \cdot \delta\mu\, u_h \tag{I.28}$$

with

$$\mathcal{E}_h^{n+1/2} = \frac{1}{2}\widetilde{m}\left(\delta u_h, \delta u_h\right) + \frac{1}{2}a\left(\mu u_h, \mu u_h\right) + \frac{1}{2}\ell\left(z_{h,1}^{n+1/2}, z_{h,1}^{n+1/2}\right) + \frac{1}{2}\ell\left(z_{h,2}^{n+1/2}, z_{h,2}^{n+1/2}\right) \tag{I.29}$$

and $\widetilde{m}(u,v) = m(u,v) + \Delta t^2\left(\theta - \frac{1}{4}\right)a(u,v)$

Proof *We use the scheme* (I.27) *with* $u_h^* = \delta\mu\, u_h \in \mathcal{Q}_h$, $z_{h,1}^* = \mu^{1/2} z_{h,1} \in \mathcal{Z}_{h,1}$ *and* $z_{h,2}^* = \mu^{1/2} z_{h,2} \in \mathcal{Z}_{h,2}$.

2.1.4 Matrix formulation

Evaluating (I.24) with basis functions as test functions allows us to compute the finite elements matrices:

$$\begin{cases}
\forall (i,j) \in [\![1, n_h^u]\!]^2, \quad (M_h)_{i,j} = m(\varphi_j, \varphi_i) = \oint_\Omega M\varphi_j \cdot \varphi_i & \text{(I.30a)}\\
\forall (i,j) \in [\![1, n_h^u]\!]^2, \quad (K_h)_{i,j} = a(\varphi_j, \varphi_i) = \oint_\Omega A\nabla\varphi_j : \nabla\varphi_i & \text{(I.30b)}\\
\forall (i,j) \in [\![1, n_h^{z1}]\!]^2, \quad (L_{h,1})_{i,j} = \ell(\phi_j, \phi_i) = \oint_\Omega \phi_j \cdot \phi_i & \text{(I.30c)}\\
\forall (i,j) \in [\![1, n_h^{z2}]\!]^2, \quad (L_{h,2})_{i,j} = \ell(\psi_j, \psi_i) = \oint_\Omega \psi_j \cdot \psi_i & \text{(I.30d)}\\
\forall (i,j) \in [\![1, n_h^{z1}]\!] \times [\![1, n_h^u]\!], \quad (\mathbb{G}_1(U_h))_{i,j} = \bar{G}_1(u_h, \varphi_j, \phi_i) = \oint_\Omega \phi_i G_1(u_h) \cdot \varphi_j & \text{(I.30e)}\\
\forall (i,j) \in [\![1, n_h^{z2}]\!] \times [\![1, n_h^u]\!], \quad (\mathbb{G}_2(U_h))_{i,j} = \bar{G}_2(u_h, \varphi_j, \psi_i) = \oint_\Omega \psi_i G_2(\nabla u_h) : \nabla\varphi_j & \text{(I.30f)}\\
\forall i \in [\![1, n_h^u]\!], \quad (F_h)_i = f(\varphi_i) = \oint_\Omega f_h \cdot \varphi_i & \text{(I.30g)}
\end{cases}$$

so that the matrix formulation of the system writes:

$$\begin{cases}
M_h\ddot{U}_h + K_h U_h + {}^t\mathbb{G}_2(U_h)Z_{h,2} + {}^t\mathbb{G}_1(U_h)Z_{h,1} = F_h & \text{(I.31a)}\\
L_{h,1}\dot{Z}_{h,1} = \mathbb{G}_1(U_h)\dot{U}_h & \text{(I.31b)}\\
L_{h,2}\dot{Z}_{h,2} = \mathbb{G}_2(U_h)\dot{U}_h & \text{(I.31c)}
\end{cases}$$

The time scheme rewrites:

$$\begin{cases}
M_h\delta^2 U_h + K_h\left\{U_h\right\}_\theta^n + {}^t\mathbb{G}_1(U_h^n)\mu^{1/2}Z_{h,1} + {}^t\mathbb{G}_2(U_h^n)\mu^{1/2}Z_{h,2} = F_h^n & \text{(I.32a)}\\
L_{h,1}\delta^{1/2}Z_{h,1} = \mathbb{G}_1(U_h^n)\delta\mu\, U_h & \text{(I.32b)}\\
L_{h,2}\delta^{1/2}Z_{h,2} = \mathbb{G}_2(U_h^n)\delta\mu\, U_h & \text{(I.32c)}
\end{cases}$$

2.1.5 Practical solution of the scheme

The numerical scheme above is linearly implicit so it can be solved as a linear system. There are two main ways to solve it though.

We can solve it with the coupled unknowns at every step:

$$\begin{pmatrix} \frac{M_h}{\Delta t^2} + \theta K_h & \frac{1}{2}\,{}^t\mathbb{G}_1(U_h^n) & \frac{1}{2}\,{}^t\mathbb{G}_2(U_h^n) \\ -\frac{1}{2}\mathbb{G}_1(U_h^n) & L_{h,1} & 0 \\ -\frac{1}{2}\mathbb{G}_2(U_h^n) & 0 & L_{h,2} \end{pmatrix} \begin{pmatrix} U_h^{n+1} \\ Z_{h,1}^{n+1/2} \\ Z_{h,2}^{n+1/2} \end{pmatrix}$$
$$= \begin{pmatrix} F_h^n + M_h \frac{2U_h^n - U_h^{n-1}}{\Delta t^2} + K_h\left((2\theta-1)U_h^n - \theta U_h^{n-1}\right) - \frac{1}{2}\,{}^t\mathbb{G}_1(U_h^n) Z_{h,1}^{n-1/2} - \frac{1}{2}\,{}^t\mathbb{G}_2(U_h^n) Z_{h,2}^{n-1/2} \\ L_{h,1} Z_{h,1}^{n-1/2} - \frac{1}{2}\mathbb{G}_1(U_h^n) U_h^{n-1} \\ L_{h,2} Z_{h,2}^{n-1/2} - \frac{1}{2}\mathbb{G}_2(U_h^n) U_h^{n-1} \end{pmatrix} \tag{I.33}$$

Or since it is linearly implicit we can eliminate $Z_{h,1}$ and $Z_{h,2}$ from (I.32a) with the relation:

$$\mu^{1/2} Z_h = \frac{\Delta t}{2}\,\delta^{1/2} Z_h + Z_h^{n-1/2} = \frac{\Delta t}{2}\,L_h^{-1}\mathbb{G}(U_h^n)\delta\mu\, U_h + Z_h^{n-1/2} \tag{I.34}$$

which leads to

$$\left\{ \begin{aligned} &\left[\frac{M_h}{\Delta t^2} + \theta K_h + \frac{1}{4}\,{}^t\mathbb{G}_1(U_h^n) L_{h,1}^{-1}\mathbb{G}_1(U_h^n) + \frac{1}{4}\,{}^t\mathbb{G}_2(U_h^n) L_{h,2}^{-1}\mathbb{G}_2(U_h^n)\right] U_h^{n+1} \\ &\quad = F_h^n + M_h \frac{2U_h^n - U_h^{n-1}}{\Delta t^2} + K_h\left((2\theta-1)U_h^n - \theta U_h^{n-1}\right) \\ &\qquad\qquad - {}^t\mathbb{G}_1(U_h^n) Z_{h,1}^{n-1/2} + \frac{1}{4}\,{}^t\mathbb{G}_1(U_h^n) L_{h,1}^{-1}\mathbb{G}_1(U_h^n) U_h^{n-1} \\ &\qquad\qquad - {}^t\mathbb{G}_2(U_h^n) Z_{h,2}^{n-1/2} + \frac{1}{4}\,{}^t\mathbb{G}_2(U_h^n) L_{h,2}^{-1}\mathbb{G}_2(U_h^n) U_h^{n-1} && \text{(I.35a)} \\ &Z_{h,1}^{n+1/2} = Z_{h,1}^{n-1/2} + L_{h,1}^{-1}\mathbb{G}_1(U_h^n)\frac{U_h^{n+1} - U_h^{n-1}}{2} && \text{(I.35b)} \\ &Z_{h,2}^{n+1/2} = Z_{h,2}^{n-1/2} + L_{h,2}^{-1}\mathbb{G}_2(U_h^n)\frac{U_h^{n+1} - U_h^{n-1}}{2} && \text{(I.35c)} \end{aligned} \right.$$

We compute U_h^{n+1} with the first equation and then compute $Z_{h,1}^{n+1/2}$ and $Z_{h,2}^{n+1/2}$ explicitly with the others.

The second formulation with elimination reduces the size of the linear system to solve by $n_h^{z1} + n_h^{z2}$ and we can use Woodbury matrix inversion formula:

Lemma 2.1 (Woodbury inversion formula.) *[Woodbury, 1950]*
Let A be a invertible square matrix of size $n \in \mathbb{N}^*$ and C one of size $p \in \mathbb{N}^*$.
U, V two matrices of size $n \times p$ and $p \times n$.
$A + UCV$ is invertible if and only if $\det\left(C^{-1} + VA^{-1}U\right) \neq 0$ and we have:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1} VA^{-1}$$

with $A = \dfrac{M_h}{\Delta t^2} + \theta K_h$, $C = \begin{pmatrix} L_{h,1}^{-1} & 0 \\ 0 & L_{h,2}^{-1} \end{pmatrix}$ and ${}^tU = V = \frac{1}{2}\begin{pmatrix} \mathbb{G}_1(U_h^n) \\ \mathbb{G}_2(U_h^n) \end{pmatrix}$.
In this case the size $p = n_h^{z1} + n_h^{z2}$ of the matrices U and V is quite large so the inversion of $C^{-1} + VA^{-1}U$ is expensive. We will see that the SAV method allows to reduce considerably the size p.

2.2 Scalar Auxiliary Variable (SAV)

Assumption 2.4 *Let u be the solution of I.3.*
We assume that there exist $(\beta_1, \beta_2) \in (\mathbb{R}_+^*)^2$ depending on the source f such that for all $(x,t) \in \Omega \times [0,T]$ $-\frac{\beta_1}{2} < \int_\Omega F_1(u(x,t))dx$ and $-\frac{\beta_2}{2} < \int_\Omega F_2(\nabla u(x,t))dx$.

It ensures that $2\int_\Omega F_1(u(x,t))dx + c_1 > c_1 - \beta_1 > 0$ and $2\int_\Omega F_2(\nabla u(x,t))dx + c_2 > c_2 - \beta_2 > 0$

Similarly to IEQ, we introduce two auxiliary variable. But now the nonlinear function is integrated on the space domain so these no longer depend on the space variable x:

$$\forall t \in [0,T], \quad \begin{cases} z_1(t) \equiv \sqrt{2\displaystyle\int_\Omega F_1(u(x,t)) + c_1} & \text{(I.36a)} \\ z_2(t) \equiv \sqrt{2\displaystyle\int_\Omega F_2(\nabla u(x,t)) + c_2} & \text{(I.36b)} \end{cases}$$

The two auxiliary functions are:

$$\begin{cases} \forall u \in H^1(\Omega)^p, \forall (x,t) \in \Omega \times [0,T], \quad g_1(u(x,t)) = \dfrac{1}{\sqrt{2\int_\Omega F_1(u(s,t))ds + c_1}} \nabla F_1(u(x,t)) & \text{(I.37a)} \\ \forall u \in H^1(\Omega)^p, \forall (x,t) \in \Omega \times [0,T], \quad g_2(u(x,t)) = \dfrac{1}{\sqrt{2\int_\Omega F_2(\nabla u(s,t))ds + c_2}} \nabla F_2(\nabla u(x,t)) & \text{(I.37b)} \end{cases}$$

but unlike IEQ these functions cannot be studied on $\mathbb{R}^p$ independently of the field u because of the integration in the square root.

The two associated operators write

$$\begin{cases} \forall u \in H^1(\Omega)^p, \quad G_1(u) = \dfrac{1}{\sqrt{2\int_\Omega F_1(u) + c_1}} \nabla F_1(u) & \text{(I.38a)} \\ \forall u \in H^1(\Omega)^p, \quad G_2(u) = \dfrac{1}{\sqrt{2\int_\Omega F_2(\nabla u) + c_2}} \nabla F_2(\nabla u) & \text{(I.38b)} \end{cases}$$

so that for $(x,t) \in \Omega \times [0,T]$, $G_i(u)(x,t) = g_i(u(x,t))$.

The nonlinear forms involved in the variational formulations are called

$$\begin{cases} \forall (u,v) \in H^1(\Omega)^p \times H^1(\Omega)^p, \quad \bar{G}_1(u,v) = \displaystyle\int_\Omega G_1(u) \cdot v & \text{(I.39a)} \\ \forall (u,v) \in H^1(\Omega)^p \times H^1(\Omega)^p, \quad \bar{G}_2(u,v) = \displaystyle\int_\Omega G_2(u) : v & \text{(I.39b)} \end{cases}$$

2.2.1 Variational formulation and Energy

The new system is derivated directly from the variational formulation (I.3):

Proposition 2.3 (Weak formulation of SAV)
Find $u \in \left(H^1(\Omega)\right)^p$ *and* $(z_1, z_2) \in \left(\mathbb{R}^{[0,T]}\right)^2$ *such that for all* $u^* \in \left(H^1(\Omega)\right)^p$

$$\begin{cases} m(\partial_t^2 u, u^*) + a(u, u^*) + z_2(t)\bar{G}_2(u, u^*) + z_1(t)\bar{G}_1(u, u^*) = f(u^*) & \text{(I.40a)} \\ \partial_t z_1(t) = \bar{G}_1(u, \partial_t u) & \text{(I.40b)} \\ \partial_t z_2(t) = \bar{G}_2(u, \partial_t u) & \text{(I.40c)} \end{cases}$$

Notice that the two auxiliary equations are not in a weak form since the auxiliary variables only depend on time and not on space anymore.

Theorem 2.6 (Energy of SAV formulation)
Any solution to I.40 verifies

$$\frac{d\mathcal{E}}{dt} = f(\partial_t u) = \int_\Omega f \cdot \partial_t u \tag{I.41}$$

with

$$\mathcal{E}(t) = \frac{1}{2} m(\partial_t u, \partial_t u) + \frac{1}{2} a(u, u) + \frac{1}{2} z_2^2 + \frac{1}{2} z_1^2 = \mathcal{E}_0(t) + \frac{c_1 + c_2}{2} \tag{I.42}$$

Proof *Just use the weak formulation I.40 with* $u^* = \partial_t u \in H^1(\Omega)^p$.

2.2.2 Spatial discretization

We use the same space discretisation as for IEQ formulation 2.1.2 but now only u is discretized since z_i are scalars.

Find $u_h \in \mathcal{Q}_h$ and $(z_{h,1}, z_{h,2}) \in \mathbb{R}^2$ such that for all $u_h^* \in \mathcal{Q}_h$:

$$\begin{cases} m(\partial_t^2 u_h, u_h^*) + a(u_h, u_h^*) + z_{h,2}\bar{G}_2(u_h, u_h^*) + z_{h,1}\bar{G}_1(u_h, u_h^*) = f(u_h^*) & \text{(I.43a)} \\ \partial_t z_{h,1} = \bar{G}_1(u_h, \partial_t u_h) & \text{(I.43b)} \\ \partial_t z_{h,2} = \bar{G}_2(u_h, \partial_t u_h) & \text{(I.43c)} \end{cases}$$

2.2.3 Time discretization

A very similar time scheme to IEQ I.27 is used here for SAV:

Numerical Scheme 2.2 (SAV Time Scheme)
We seek $u_h \in \mathcal{Q}_h$ *and* $(z_1, z_2) \in \left(\mathbb{R}^{[0,T]}\right)^2$ *so that for all* $u_h^* \in \mathcal{Q}_h$ *and all* $n \in [\![0, N]\!]$

$$\begin{cases} m(\delta^2 u_h, u_h^*) + a(\{u_h\}_\theta^n, u_h^*) + \mu^{1/2} z_{h,2}\, \bar{G}_2(u_h^n, u_h^*) + \mu^{1/2} z_{h,1}\, \bar{G}_1(u_h^n, u_h^*) = f^n(u_h^*) & \text{(I.44a)} \\ \delta^{1/2} z_{h,1} = \bar{G}_1(u_h^n, \delta\mu u_h) & \text{(I.44b)} \\ \delta^{1/2} z_{h,2} = \bar{G}_2(u_h^n, \delta\mu u_h) & \text{(I.44c)} \end{cases}$$

Theorem 2.8 (Discrete Energy Identity of SAV scheme)
Any solution to I.44 satisfies

$$\delta^{1/2}\mathcal{E}_h = f^n(\delta\mu u_h) = \int_\Omega f_h^n \cdot \delta\mu u_h \tag{I.45}$$

with

$$\mathcal{E}_h^{n+1/2} = \frac{1}{2}\widetilde{m}(\delta u_h, \delta u_h) + \frac{1}{2}a(\mu u_h, \mu u_h) + \frac{1}{2}\left(z_{h,1}^{n+1/2}\right)^2 + \frac{1}{2}\left(z_{h,2}^{n+1/2}\right)^2 \tag{I.46}$$

and $\widetilde{m}(u,v) = m(u,v) + \Delta t^2\left(\theta - \frac{1}{4}\right)a(u,v)$

Proof *We use the scheme* (I.44) *with* $u_h^* = \delta\mu\, u_h$.

2.2.4 Matrix formulation

We can compute the Finite Elements matrices and vectors:

$$\begin{cases} \forall (i,j) \in [\![1, n_h^u]\!]^2, \quad (M_h)_{i,j} = m(\varphi_j, \varphi_i) = \oint_\Omega M\varphi_j \cdot \varphi_i & \text{(I.47a)} \\ \forall (i,j) \in [\![1, n_h^u]\!]^2, \quad (K_h)_{i,j} = a(\varphi_j, \varphi_i) = \oint_\Omega A\nabla\varphi_j : \nabla\varphi_i & \text{(I.47b)} \\ \forall i \in [\![1, n_h^u]\!], \quad (\mathbb{G}_1(U_h))_i = \bar{G}_1(u_h, \varphi_i) & \text{(I.47c)} \\ \forall i \in [\![1, n_h^u]\!], \quad (\mathbb{G}_2(U_h))_i = \bar{G}_2(u_h, \varphi_i) & \text{(I.47d)} \\ \forall i \in [\![1, n_h^u]\!], \quad (F_h)_i = f(\varphi_i) = \oint_\Omega f_h \cdot \varphi_i & \text{(I.47e)} \end{cases}$$

so that the matrix formulation of the system (I.43) writes:

$$\begin{cases} M_h\ddot{U}_h + K_hU_h + \mathbb{G}_2(U_h)z_{h,2} + \mathbb{G}_1(U_h)z_{h,1} = F_h & \text{(I.48a)} \\ \dot{z}_{h,1} = \mathbb{G}_1(U_h)\cdot\dot{U}_h & \text{(I.48b)} \\ \dot{z}_{h,2} = \mathbb{G}_2(U_h)\cdot\dot{U}_h & \text{(I.48c)} \end{cases}$$

the time scheme rewrites:

$$\begin{cases} M_h\delta^2U_h + K_h\left\{U_h\right\}_\theta^n + \mathbb{G}_2(U_h^n)\mu^{1/2}z_{h,2} + \mathbb{G}_1(U_h^n)\mu^{1/2}z_{h,1} = F_h^n & \text{(I.49a)} \\ \delta^{1/2}z_{h,1} = \mathbb{G}_1(U_h^n)\cdot\delta\mu U_h & \text{(I.49b)} \\ \delta^{1/2}z_{h,2} = \mathbb{G}_2(U_h^n)\cdot\delta\mu U_h & \text{(I.49c)} \end{cases}$$

2.2.5 Practical solution of the scheme

The SAV schemes can be very efficiently solved. The one presented just above is not an exception. Just like in 2.1.5 we eliminate $z_{h,1}$ and $z_{h,2}$ with the relation:

$$\mu^{1/2}z_h = \frac{\Delta t}{2}\delta^{1/2}z_h + z_h^{n-1/2} = \frac{\Delta t}{2}\mathbb{G}(U_h^n)\cdot\delta\mu U_h + z_h^{n-1/2} \tag{I.50}$$

which leads to

$$\begin{cases} \left[\dfrac{M_h}{\Delta t^2} + \theta K_h + \dfrac{1}{4}\mathbb{G}_1(U_h^n)\,{}^t\mathbb{G}_1(U_h^n) + \dfrac{1}{4}\mathbb{G}_2(U_h^n)\,{}^t\mathbb{G}_2(U_h^n)\right] U_h^{n+1} \\ \qquad = F_h^n + M_h \dfrac{2U_h^n - U_h^{n-1}}{\Delta t^2} + K_h\left((2\theta-1)U_h^n - \theta U_h^{n-1}\right) \\ \qquad\qquad - \mathbb{G}_1(U_h^n) z_{h,1}^{n-1/2} + \dfrac{1}{4}\mathbb{G}_1(U_h^n)\,{}^t\mathbb{G}_1(U_h^n) U_h^{n-1} \\ \qquad\qquad - \mathbb{G}_2(U_h^n) z_{h,2}^{n-1/2} + \dfrac{1}{4}\mathbb{G}_2(U_h^n)\,{}^t\mathbb{G}_2(U_h^n) U_h^{n-1} \\ z_{h,1}^{n+1/2} = z_{h,1}^{n-1/2} + \mathbb{G}_1(U_h^n)\cdot\dfrac{U_h^{n+1} - U_h^{n-1}}{2} \\ z_{h,2}^{n+1/2} = z_{h,2}^{n-1/2} + \mathbb{G}_2(U_h^n)\cdot\dfrac{U_h^{n+1} - U_h^{n-1}}{2} \end{cases} \tag{I.51a} \tag{I.51b} \tag{I.51c}$$

and now we can use Woodbury inversion formula 2.1 to solve (I.51a) with $A = \dfrac{M_h}{\Delta t^2} + \theta K_h$, $C = I_2$ and $U = {}^tV = \frac{1}{2}\begin{pmatrix}\mathbb{G}_1(U_h^n)\\ \mathbb{G}_2(U_h^n)\end{pmatrix}$.
The size p of the matrix $I_2 + VA^{-1}U$ to invert at every time step is now only 2.

Remark 2.1 *If only one auxiliary variable is used, Woodbury identity becomes Sherman-Morrison formula and the computation is even more efficient since no matrix inversion is required.*

Lemma 2.2 (Sherman-Morrison formula.) *[Sherman and Morrison, 1950]*
Let A be a invertible square matrix of size $n \in \mathbb{N}^*$ and u, v two column vectors of size n.
$A + u\,{}^tv$ is invertible if and only if $1 + {}^tvA^{-1}u \neq 0$ and we have:

$$\left(A + u\,{}^tv\right)^{-1} = A^{-1} - \frac{A^{-1}u\,{}^tvA^{-1}}{1 + {}^tvA^{-1}u}$$

Remark 2.2 *This formula is a specific application case of Woodbury inversion formula in lemma 2.1 when the contribution added to A is of rank 1.*
The matrix A which is very often symmetric and its inverse (or at least its LU factorization) A^{-1} are computed once and for all at the beginning of the simulation since they do not depend of time. Then if b is the right-hand side of (I.51a), to implement the Sherman-Morrison formula we do at every time step:

- compute $A^{-1}b$ and $A^{-1}u$, and notice that ${}^tuA^{-1} = {}^t(A^{-1}u)$ for a symmetric matrix A
- compute $s = \frac{{}^tuA^{-1}b}{1+{}^tuA^{-1}u} = \frac{(A^{-1}u)\cdot b}{1+u\cdot(A^{-1}u)}$ which is a scalar
- and finally compute $A^{-1}b + s \times (A^{-1}u)$

We see that solution of (I.51a) only requires some matrix-vector and scalar products, but no extra heavy LU factorization at every step which is a huge gain of SAV over IEQ.

3 Numerical analysis

In this section we analyze the mathematical properties of the IEQ scheme. SAV scheme is very similar so the results adapt easily.

It is important to recall that we study two different types of nonlinear terms:

- the ones of type 1 like F_1 composed with u, $F_1(u)$
- the ones of type 2 like F_2 composed with ∇u, $F_2(\nabla u)$

A time convergence proof for type 1 nonlinear terms is given. It is uniform with respect to the CFL condition. For type 2 we only give a time convergence proof in finite dimension with fixed space discretization and emphasize the blocking elements for a uniform estimation of the errors.

3.1 Stability

In this section we will work with the scalar product induced by the bilinear form m. In other words we denote $(u,v)_m = m(u,v)$ and $\|v\|_m = \sqrt{m(v,v)}$ the associated norm.

Assumption 3.1 (Equivalence of norms)
In the following we asses that there exists two constants c_m and C_m such that

$$c_m \|\cdot\|_{(L^2(\Omega))^p} \leqslant \|\cdot\|_m \leqslant C_m \|\cdot\|_{(L^2(\Omega))^p}$$

meaning that the bilinear form m is coercive and continuous in $\left(L^2(\Omega)\right)^p$.

Let's first recall the scheme:
Seek $(u_h, z_{h,1}, z_{h,2}) \in \mathcal{Q}_h \times \mathcal{Z}_{h,1} \times \mathcal{Z}_{h,2}$ so that for all $(u_h^*, z_{h,1}^*, z_{h,2}^*) \in \mathcal{Q}_h \times \mathcal{Z}_{h,1} \times \mathcal{Z}_{h,2}$ and all $n \in [\![0, N]\!]$

$$\begin{cases} \left(\delta^2 u_h, u_h^*\right) + a\left(\{u_h\}_\theta^n, u_h^*\right) + \bar{G}_1\left(u_h^n, u_h^*, \mu^{1/2} z_{h,1}\right) + \bar{G}_2\left(u_h^n, u_h^*, \mu^{1/2} z_{h,2}\right) = f^n(u_h^*) & \text{(I.52a)} \\ \ell\left(\delta^{1/2} z_{h,1}, z_{h,1}^*\right) = \bar{G}_1\left(u_h^n, \delta\mu\, u_h, z_{h,1}^*\right) & \text{(I.52b)} \\ \ell\left(\delta^{1/2} z_{h,2}, z_{h,2}^*\right) = \bar{G}_2\left(u_h^n, \delta\mu\, u_h, z_{h,2}^*\right) & \text{(I.52c)} \end{cases}$$

and its energy identity:

$$\delta^{1/2}\mathcal{E}_h = f^n\left(\delta\mu\, u_h\right) = \int_\Omega f_h^n \cdot \delta\mu\, u_h \tag{I.53}$$

with

$$\mathcal{E}_h^{n+1/2} = \frac{1}{2}\widetilde{m}\left(\delta u_h, \delta u_h\right) + \frac{1}{2}a\left(\mu u_h, \mu u_h\right) + \frac{1}{2}\ell\left(z_{h,1}^{n+1/2}, z_{h,1}^{n+1/2}\right) + \frac{1}{2}\ell\left(z_{h,2}^{n+1/2}, z_{h,2}^{n+1/2}\right) \tag{I.54}$$

and $\quad \widetilde{m}(u,v) = (u,v)_m + \Delta t^2\left(\theta - \frac{1}{4}\right)a(u,v)$

In this section we use a technique of [Chabassier and Imperiale, 2017] to show uniform stability of the scheme with respect to a CFL condition.

Definition 3.1
For each space step h we define an operator $A_h : \mathcal{Q}_h \longrightarrow \mathcal{Q}_h$ such that

$$\forall v_h \in \mathcal{Q}_h, \quad a(u_h, v_h) = (A_h u_h, v_h)_m \tag{I.55}$$

Assumption 3.2
The operator A_h is self-adjoint and positive and there exists $(c_0, c_a) \in \left(\mathbb{R}_+^*\right)^2$ such that

$$\begin{cases} \forall v_h \in \mathcal{Q}_h, \quad a(v_h, v_h) \geqslant 0 & \text{(I.56a)} \\ \forall v_h \in \mathcal{Q}_h, \quad a(v_h, v_h) + c_0 \left\|v_h\right\|_m^2 \geqslant c_a \left\|v_h\right\|_{(H^1(\Omega))^p}^2 & \text{(I.56b)} \end{cases}$$

Proposition 3.1
The spectral theorem applies to A_h. We denote by $Sp(A_h)$ its spectrum which is a finite set of positive eigenvalues, and there exist a family $\left(w_h^\lambda\right)_{\lambda \in Sp(A_h)}$ which is a orthonormal basis of $\mathcal{Q}_h$ such that

$$\forall v_h \in \mathcal{Q}_h, \quad \left(A_h w_h^\lambda, v_h\right)_m = \lambda \left(w_h^\lambda, v_h\right)_m \tag{I.57}$$

and $\left\|w_h^\lambda\right\|_m = 1$

The energy I.54 can be written with two polynomials $\mathcal{P}_K$ and $\mathcal{P}_P$ such that:

$$\begin{cases} \mathcal{E}_h^{n+1/2} = \dfrac{1}{2}\left(\mathcal{P}_K\left(\Delta t^2 A_h\right)\delta u_h, \delta u_h\right) + \dfrac{1}{2}\left(A_h \mathcal{P}_P\left(\Delta t^2 A_h\right)\mu u_h, \mu u_h\right) \\ \qquad\qquad + \dfrac{1}{2}\ell\left(z_{h,1}^{n+1/2}, z_{h,1}^{n+1/2}\right) + \dfrac{1}{2}\ell\left(z_{h,2}^{n+1/2}, z_{h,2}^{n+1/2}\right) & \text{(I.58a)} \\ \mathcal{P}_K(X) = 1 + \left(\theta - \dfrac{1}{4}\right) X & \text{(I.58b)} \\ \mathcal{P}_P(X) = 1 & \text{(I.58c)} \end{cases}$$

Proposition 3.2 (CFL Condition)
The energy $\mathcal{E}_h^{n+1/2}$ is positive for all steps n if $\mathcal{P}_K(x) \geqslant 0$ and $x\mathcal{P}_P(x) \geqslant 0$ for all $x \in [0, \eta]$:

- *If $\theta \geqslant \frac{1}{4}$, $\mathcal{E}_h^{n+1/2}$ is always positive and $\eta = +\infty$*
- *If $\theta < \frac{1}{4}$ it is positive with CFL condition:*

$$\Delta t^2 \rho\left(A_h\right) \leqslant \eta = \frac{4}{1 - 4\theta} \tag{I.59}$$

where $\rho\left(A_h\right)$ is the largest eigenvalue of A_h.

Definition 3.2
We suppose that there exists a a partition of $[0, \eta] = J_K \cup J_P$ such that there exists two real constants $C_K > 0$ and $C_P > 0$ such that

$$\forall x \in J_K, \quad \mathcal{P}_K(x) \geqslant C_K \qquad \text{and} \qquad \forall x \in J_P, \quad x\mathcal{P}_P(x) \geqslant C_P \tag{I.60}$$

and we introduce two projectors on corresponding frequencies of A_h:

$$\forall v_h \in \mathcal{Q}_h, \quad \begin{cases} \Pi_K v_h = \displaystyle\sum_{\substack{\Delta t^2 \lambda \in J_K \\ \lambda \in Sp(A_h)}} \left(w_h^\lambda, v_h\right)_m w_h^\lambda & \text{(I.61a)} \\ \Pi_P v_h = \displaystyle\sum_{\substack{\Delta t^2 \lambda \in J_P \\ \lambda \in Sp(A_h)}} \left(w_h^\lambda, v_h\right)_m w_h^\lambda & \text{(I.61b)} \end{cases}$$

Theorem 3.1 (Stability)
The IEQ scheme I.27 and SAV scheme I.44 are stable if the CFL condition 3.2 is satisfied.

$$\begin{cases} \sqrt{\mathcal{E}_h^{n+1/2}} \leqslant \sqrt{\mathcal{E}_h^{1/2}} + \gamma\sqrt{2}\Delta t \sum_{j=1}^{n} \left\| f_h^j \right\|_{(L^2(\Omega))^p} & \text{(I.62a)} \\ \left\| u_h^{n+1} \right\|_{(L^2(\Omega))^p} \leqslant C_m\sqrt{2} \left\| u_h^0 \right\|_{(L^2(\Omega))^p} + 2\gamma t^n \sqrt{2\mathcal{E}_h^{1/2}} + 4\gamma^2 \Delta t^2 \sum_{i=0}^{n}\sum_{j=1}^{i} \left\| f_h^j \right\|_{(L^2(\Omega))^p} & \text{(I.62b)} \\ \left\| \delta\mu\, u_h \right\|_{(L^2(\Omega))^p} \leqslant 2\gamma\sqrt{2\mathcal{E}_h^{1/2}} + 4\gamma^2\Delta t \sum_{j=1}^{n} \left\| f_h^j \right\|_{(L^2(\Omega))^p} & \text{(I.62c)} \\ \left\| z_{h,i}^{n+1/2} \right\|_{L^2(\Omega)} \leqslant \sqrt{2\mathcal{E}_h^{1/2}} + 2\gamma\Delta t \sum_{j=1}^{n} \left\| f_h^j \right\|_{(L^2(\Omega))^p} \quad \textit{for IEQ scheme} & \text{(I.62d)} \\ \left| z_{h,i}^{n+1/2} \right| \leqslant \sqrt{2\mathcal{E}_h^{1/2}} + 2\gamma\Delta t \sum_{j=1}^{n} \left\| f_h^j \right\|_{(L^2(\Omega))^p} \quad \textit{for SAV scheme} & \text{(I.62e)} \end{cases}$$

with $\gamma = \frac{C_P^{-1/2} + \frac{1}{2}C_K^{-1/2}}{c_m}$.

Proof
The following proof is given in [Chabassier and Imperiale, 2017] for various linear schemes. It is adapted here for nonlinear schemes.

First notice that I.60 implies that for u_h *solution of I.27*

$$\begin{cases} \left\| \Pi_K \delta u_h \right\|_m^2 \leqslant C_K^{-1} \left(\mathcal{P}_K \left(\Delta t^2 A_h \right) \delta u_h, \delta u_h \right)_m \leqslant 2C_K^{-1} \mathcal{E}_h^{n+1/2} & \text{(I.63a)} \\ \left\| \Pi_P \mu u_h \right\|_m^2 \leqslant C_P^{-1} \left(\Delta t^2 A_h \mathcal{P}_P \left(\Delta t^2 A_h \right) \mu u_h, \mu u_h \right)_m \leqslant 2C_P^{-1} \Delta t^2 \mathcal{E}_h^{n+1/2} & \text{(I.63b)} \end{cases}$$

We start with an estimate on $\left\| \delta\mu\, u_h \right\|_m$*:*

$$\begin{aligned} c_m \left\| \delta\mu\, u_h \right\|_{(L^2(\Omega))^p} \leqslant \left\| \delta\mu\, u_h \right\|_m &\leqslant \left\| \Pi_K \delta\mu\, u_h \right\|_m + \left\| \Pi_P \delta\mu\, u_h \right\|_m & \text{(I.64)} \\ &\leqslant \mu \left\| \Pi_K \delta u_h \right\|_m + \frac{2}{\Delta t}\mu \left\| \Pi_P \mu u_h \right\|_m & \text{(I.65)} \\ &\leqslant \mu \left[C_K^{-1/2} \sqrt{2\mathcal{E}_h} \right] + \frac{2}{\Delta t}\mu \left[\Delta t C_P^{-1/2} \sqrt{2\mathcal{E}_h} \right] & \text{(I.66)} \\ &\leqslant c_m \gamma\sqrt{2} \left[\sqrt{\mathcal{E}_h^{n+1/2}} + \sqrt{\mathcal{E}_h^{n-1/2}} \right] & \text{(I.67)} \end{aligned}$$

Now we can apply Cauchy-Schwarz to (I.53)*:*

$$\frac{1}{\Delta t}\left(\mathcal{E}_h^{n+1/2} - \mathcal{E}_h^{n-1/2} \right) \leqslant \left\| f_h^n \right\|_{(L^2(\Omega))^p} \left\| \delta\mu\, u_h \right\|_{(L^2(\Omega))^p} \leqslant \gamma\sqrt{2} \left\| f_h^n \right\|_{(L^2(\Omega))^p} \left[\sqrt{\mathcal{E}_h^{n+1/2}} + \sqrt{\mathcal{E}_h^{n-1/2}} \right] \quad \text{(I.68)}$$

We can simplify and sum from 1 to n which gives the first result I.62a with telescopic sum.

For the second estimate on $\left\| u_h^{n+1} \right\|_{(L^2(\Omega))^p}$ *we write:*

$$\begin{aligned} \left\| \Pi_K u_h^{n+1} \right\|_m &\leqslant \left\| \Pi_K u_h^n \right\|_m + \Delta t \left\| \Pi_K \delta u_h \right\|_m & \text{(I.69)} \\ &\leqslant \left\| \Pi_K u_h^n \right\|_m + \Delta t C_K^{-1/2} \sqrt{2\mathcal{E}_h^{n+1/2}} & \text{(I.70)} \end{aligned}$$

and

$$\begin{aligned} \left\| \Pi_P u_h^{n+1} \right\|_m &\leqslant \left\| \Pi_P u_h^n \right\|_m + 2 \left\| \Pi_P \mu u_h \right\|_m & \text{(I.71)} \\ &\leqslant \left\| \Pi_P u_h^n \right\|_m + 2\Delta t C_K^{-1/2} \sqrt{2\mathcal{E}_h^{n+1/2}} & \text{(I.72)} \end{aligned}$$

which implies with telescopic sums that

$$c_m \left\|u_h^{n+1}\right\|_{(L^2(\Omega))^p} \leqslant \left\|u_h^{n+1}\right\|_m \leqslant \left\|\Pi_K u_h^{n+1}\right\|_m + \left\|\Pi_P u_h^{n+1}\right\|_m \leqslant \sqrt{2}\left\|u_h^0\right\|_m + 2c_m\gamma\sqrt{2}\Delta t \sum_{i=0}^{n} \sqrt{\mathcal{E}_h^{i+1/2}} \quad \text{(I.73)}$$

and using (I.62a) *gives the result I.62b.*

The estimates on the auxiliary variables come directly from the energy:

$$\begin{cases} \left\|z_{h,i}^{n+1/2}\right\|_{L^2(\Omega)}^2 = \ell\left(z_{h,i}^{n+1/2}, z_{h,i}^{n+1/2}\right) \leqslant 2\mathcal{E}_h^{n+1/2} & \textit{for IEQ scheme} \quad \text{(I.74a)} \\ \left|z_{h,i}^{n+1/2}\right|^2 = \left(z_{h,i}^{n+1/2}\right)^2 \leqslant 2\mathcal{E}_h^{n+1/2} & \textit{for SAV scheme} \quad \text{(I.74b)} \end{cases}$$

and we just use I.62a again.

3.2 Discrete regularity assumptions

Assumption 3.3
Let $(u_h, z_{h,1}, z_{h,2})$ the solution of the semi-discrete problem I.24.
We assume that there exists a constant $C_{p,q}$ such that for all $h > 0$ we have

$$\begin{cases} \sup_{t\in[0,T]} \left\|A_h^{p/2}\partial_t^q u_h\right\|_{(L^2(\Omega))^p} \leqslant C_{p,q} & \text{(I.75a)} \\ \sup_{t\in[0,T]} \left\|\partial_t^\ell z_{h,i}\right\|_{L^2(\Omega)} \leqslant C_{zi,\ell} & \text{(I.75b)} \end{cases}$$

In particular we suppose that I.75a is true for $(p,q) \in \{(0,0),(0,3),(0,4),(1,0),(1,3),(2,2)\}$ and that I.75b is true for $\ell \in \{2,3\}$.

Let $\mathcal{J}_h = \bigcup_{n=0}^{N} \mathrm{Im}\left(u_h(t^n)\right)$ and $\mathcal{J}_{\nabla,h} = \bigcup_{n=0}^{N} \mathrm{Im}\left(\nabla u_h(t^n)\right)$

Let $\mathcal{I}_h = \mathcal{J}_h \cup \bigcup_{n=0}^{N} \mathrm{Im}\left(u_h^n\right)$ and $\mathcal{I}_{\nabla,h} = \mathcal{J}_{\nabla,h} \cup \bigcup_{n=0}^{N} \mathrm{Im}\left(\nabla u_h^n\right)$

with $\mathrm{Im}(u)$ the set of values taken by the field u.

Assumption 3.4
If the scheme is stable and for reasonably small source term and initial conditions we can assess that $\mathcal{I}_h \subset \mathcal{I}$ and $\mathcal{I}_{\nabla,h} \subset \mathcal{I}_\nabla$ with $\mathcal{I}$ and $\mathcal{I}_\nabla$ defined in assumption 2.1.

Assumption 3.5
We assume that there exist $(\beta_{h,1}, \beta_{h,2}) \in (\mathbb{R}_+^*)^2$ depending on the source f such that for all $(a,b) \in \mathcal{I}_h \times \mathcal{I}_{\nabla,h}$ $-\frac{\beta_{h,1}}{2} < F_1(a)$ and $-\frac{\beta_{h,2}}{2} < F_2(b)$.

Proposition 3.3 (Lipschitz properties of $\bar{G}$ IEQ functions)
For all $u_h^* \in \mathcal{Q}_h$ and $z_{h,1}^* \in \mathcal{Z}_{h,1}$

$$\left|\bar{G}_1\left(u_h(t^n), u_h^*, z_{h,1}^*\right) - \bar{G}_1\left(u_h^n, u_h^*, z_{h,1}^*\right)\right| \leqslant C_{g1} \left\|u_h(t^n) - u_h^n\right\|_{(L^2(\Omega))^p} \left\|u_h^*\right\|_{(L^2(\Omega))^p} \left\|z_{h,1}^*\right\|_{L^2(\Omega)} \tag{I.76}$$

with $C_{g1} = \frac{1}{c_1 - \beta_{h,1}}\left[C_{df1} \sup_{p\in\mathcal{J}_h} \sqrt{2F_1(p) + c_1} + \frac{C_{f1}}{\sqrt{c_1-\beta_{h,1}}} \sup_{p\in\mathcal{J}_h} \left\|\nabla F_1(p)\right\|_2\right]$ *and* $\beta_{h,1} = -2 \inf_{p\in\mathcal{I}_h} F_1(p)$

For all $u_h^* \in \mathcal{Q}_h$ and $z_{h,2}^* \in \mathcal{Z}_{h,2}$

$$\left|\bar{G}_2\left(u_h(t^n), u_h^*, z_{h,2}^*\right) - \bar{G}_2\left(u_h^n, u_h^*, z_{h,2}^*\right)\right| \leqslant C_{g2} \left\|u_h(t^n) - u_h^n\right\|_{(H^1(\Omega))^p} \left\|u_h^*\right\|_{(H^1(\Omega))^p} \left\|z_{h,2}^*\right\|_{L^2(\Omega)} \tag{I.77}$$

with $C_{g2} = \frac{1}{c_2 - \beta_{h,2}}\left[C_{df2} \sup_{p\in\mathcal{J}_{\nabla,h}} \sqrt{2F_2(p) + c_2} + \frac{C_{f2}}{\sqrt{c_2-\beta_{h,2}}} \sup_{p\in\mathcal{J}_{\nabla,h}} \left\|\nabla F_2(p)\right\|_2\right]$ *and* $\beta_{h,2} = -2 \inf_{p\in\mathcal{I}_{\nabla,h}} F_1(p)$

Proof *$\bar{G}_i$ functions inherit of the Lipschitz properties of functions F_i of assumption 2.1.*

$$\begin{aligned}
\left|\bar{G}_1\left(u_h(t^n), u_h^*, z_{h,1}^*\right) - \bar{G}_1\left(u_h^n, u_h^*, z_{h,1}^*\right)\right| &= \left|\int_\Omega z_{h,1}^* \left(G_1(u_h(t^n)) - G_1(u_h^n)\right) \cdot u_h^*\right| && \text{(I.78)}\\
&\leqslant \left\|G_1(u_h(t^n)) - G_1(u_h^n)\right\|_{(L^2(\Omega))^p} \left\|u_h^*\right\|_{(L^2(\Omega))^p} \left\|z_{h,1}^*\right\|_{L^2(\Omega)} && \text{(I.79)}
\end{aligned}$$

Because of the good regularity of the square root and the Lipschitz assumptions 2.1 and 3.5 on F_1 we can prove (see [Yang and Zhang, 2020]) that for all $(a,b) \in \mathcal{I}_h^2 \subset \mathcal{I}^2$:

$$\begin{aligned}
\left|\sqrt{2F_1(a) + c_1} - \sqrt{2F_1(b) + c_1}\right| &\leqslant \frac{1}{\sqrt{c_1 - \beta_{h,1}}} \left|F_1(a) - F_1(b)\right| && \text{(I.80)}\\
&\leqslant \frac{C_{f1}}{\sqrt{c_1 - \beta_{h,1}}} \left\|a - b\right\|_2 && \text{(I.81)}
\end{aligned}$$

Then we have for all $(a,b) \in \mathcal{I}_h^2 \subset \mathcal{I}^2$:

$$\begin{aligned}
\left\|g_1(a) - g_1(b)\right\|_2 &= \left\|\frac{\nabla F_1(a)}{\sqrt{2F_1(a) + c_1}} - \frac{\nabla F_1(b)}{\sqrt{2F_1(b) + c_1}}\right\|_2 && \text{(I.82)}\\
&= \left\|\frac{\sqrt{2F_1(b) + c_1}\nabla F_1(a) - \sqrt{2F_1(a) + c_1}\nabla F_1(b)}{\sqrt{2F_1(a) + c_1}\sqrt{2F_1(b) + c_1}}\right\|_2 && \text{(I.83)}\\
&\leqslant \frac{1}{c_1 - \beta_{h,1}} \left\|\sqrt{2F_1(b) + c_1}\nabla F_1(a) - \sqrt{2F_1(a) + c_1}\nabla F_1(b)\right\|_2 && \text{(I.84)}\\
&\leqslant \frac{1}{c_1 - \beta_{h,1}} \Big\|\sqrt{2F_1(b) + c_1}\left(\nabla F_1(a) - \nabla F_1(b)\right) && \text{(I.85)}\\
&\qquad + \left(\sqrt{2F_1(b) + c_1} - \sqrt{2F_1(a) + c_1}\right)\nabla F_1(b)\Big\|_2 \\
&\leqslant \frac{1}{c_1 - \beta_{h,1}}\left[C_{df1}\sqrt{2F_1(b) + c_1}\left\|a - b\right\|_2 + \frac{C_{f1}}{\sqrt{c_1 - \beta_1}}\left\|a - b\right\|_2 \left\|\nabla F_1(b)\right\|_2\right] && \text{(I.86)}\\
&\leqslant \frac{1}{c_1 - \beta_{h,1}}\left[C_{df1}\sqrt{2F_1(b) + c_1} + \frac{C_{f1}}{\sqrt{c_1 - \beta_{h,1}}}\left\|\nabla F_1(b)\right\|_2\right]\left\|a - b\right\|_2 && \text{(I.87)}\\
&\leqslant \underbrace{\frac{1}{c_1 - \beta_{h,1}}\left[C_{df1} \sup_{p\in\mathcal{J}_h} \sqrt{2F_1(p) + c_1} + \frac{C_{f1}}{\sqrt{c_1 - \beta_{h,1}}} \sup_{p\in\mathcal{J}_h} \left\|\nabla F_1(p)\right\|_2\right]}_{C_{g1}}\left\|a - b\right\|_2 && \text{(I.88)}
\end{aligned}$$

So in the end

$$\begin{aligned}
\|G_1(u_h(t^n)) - G_1(u_h^n)\|^2_{(L^2(\Omega))^p} &= \int_\Omega (G_1(u_h(t^n)) - G_1(u_h^n)) \cdot (G_1(u_h(t^n)) - G_1(u_h^n)) && \text{(I.89)}\\
&= \int_\Omega \|g_1(u_h(x,t^n)) - g_1(u_h^n(x))\|_2^2 \, dx && \text{(I.90)}\\
&\leqslant C_{g1}^2 \int_\Omega \|u_h(x,t^n) - u_h^n(x)\|_2^2 \, dx && \text{(I.91)}\\
&\leqslant C_{g1}^2 \, \|u_h(t^n) - u_h^n\|^2_{(L^2(\Omega))^p} && \text{(I.92)}
\end{aligned}$$

which gives the expected result. A similar justification applies for $\bar{G}_2$.

Assumption 3.6
We assume that there exist $(\beta_{h,1}, \beta_{h,2}) \in (\mathbb{R}_+^*)^2$ depending on the source f such that for all step n
$-\frac{\beta_{h,1}}{2} < \int_\Omega F_1(u_h^n(x))dx$ *and* $-\frac{\beta_{h,2}}{2} < \int_\Omega F_2(\nabla u_h^n(x))dx$
$-\frac{\beta_{h,1}}{2} < \int_\Omega F_1(u_h(x,t^n))dx$ *and* $-\frac{\beta_{h,2}}{2} < \int_\Omega F_2(\nabla u_h(x,t^n))dx$.

Proposition 3.4 (Lipschitz properties of $\bar{G}$ SAV functions)
For all $u_h^* \in \mathcal{Q}_h$

$$\left|\bar{G}_1\left(u_h(t^n), u_h^*\right) - \bar{G}_1\left(u_h^n, u_h^*\right)\right| \leqslant C_{g1} \, \|u_h(t^n) - u_h^n\|_{(L^2(\Omega))^p} \, \|u_h^*\|_{(L^2(\Omega))^p} \tag{I.93}$$

with $C_{g1} = \frac{1}{c_1 - \beta_{h,1}} \left[C_{df1} \sup_{n \in [\![0,N]\!]} \sqrt{2 \int_\Omega F_1(u_h(x,t^n))dx + c_1} + \frac{C_{f1}}{\sqrt{c_1 - \beta_{h,1}}} \sup_{p \in \mathcal{J}_h} \|\nabla F_1(p)\|_2 \right]$
and $\beta_{h,1} = -2 \inf_{n \in [\![0,N]\!]} \int_\Omega F_1(u_h(x,t^n))dx$

For all $u_h^* \in \mathcal{Q}_h$

$$\left|\bar{G}_2\left(u_h(t^n), u_h^*\right) - \bar{G}_2\left(u_h^n, u_h^*\right)\right| \leqslant C_{g2} \, \|u_h(t^n) - u_h^n\|_{(H^1(\Omega))^p} \, \|u_h^*\|_{(H^1(\Omega))^p} \tag{I.94}$$

with $C_{g2} = \frac{1}{c_2 - \beta_{h,2}} \left[C_{df1} \sup_{n \in [\![0,N]\!]} \sqrt{2 \int_\Omega F_2(\nabla u_h(x,t^n))dx + c_2} + \frac{C_{f1}}{\sqrt{c_2 - \beta_{h,2}}} \sup_{p \in \mathcal{J}_{\nabla,h}} \|\nabla F_2(p)\|_2 \right]$
and $\beta_{h,2} = -2 \inf_{n \in [\![0,N]\!]} \int_\Omega F_2(\nabla u_h(x,t^n))dx$

Proof *$\bar{G}_i$ functions inherit of the Lipschitz properties of functions F_i of assumption 2.1.*

$$\begin{aligned}
\left|\bar{G}_1\left(u_h(t^n), u_h^*\right) - \bar{G}_1\left(u_h^n, u_h^*\right)\right| &= \left| \int_\Omega (G_1(u_h(t^n)) - G_1(u_h^n)) \cdot u_h^* \right| && \text{(I.95)}\\
&\leqslant \|G_1(u_h(t^n)) - G_1(u_h^n)\|_{(L^2(\Omega))^p} \, \|u_h^*\|_{(L^2(\Omega))^p} && \text{(I.96)}
\end{aligned}$$

Because of the good regularity of the square root and the Lipschitz assumptions 2.1 and 3.6 we can prove that

the quantity $\mathcal{I}F_1(u) = \sqrt{2\int_\Omega F_1(u(x,t))dx + c_1}$ *verifies:*

$$
\begin{aligned}
|\mathcal{I}F_1(u_h(t^n)) - \mathcal{I}F_1(u_h^n)| &\leqslant \frac{1}{\sqrt{c_1-\beta_{h,1}}}\left|\int_\Omega F_1(u_h(x,t^n))dx - \int_\Omega F_1(u_h^n(x))dx\right| && \text{(I.97)}\\
&\leqslant \frac{1}{\sqrt{c_1-\beta_{h,1}}}\int_\Omega |F_1(u_h(x,t^n)) - F_1(u_h^n(x))|\,dx && \text{(I.98)}\\
&\leqslant \frac{C_{f1}}{\sqrt{c_1-\beta_{h,1}}}\int_\Omega \|u_h(x,t^n) - u_h^n(x)\|_2\,dx && \text{(I.99)}\\
&\leqslant \frac{C_{f1}}{\sqrt{c_1-\beta_{h,1}}}\|u_h(t^n) - u_h^n\|_{(L^1(\Omega))^p} && \text{(I.100)}\\
&\leqslant C_{f1}\sqrt{\frac{|\Omega|}{c_1-\beta_{h,1}}}\|u_h(t^n) - u_h^n\|_{(L^2(\Omega))^p} \quad \textit{thanks to Hölder inequality} && \text{(I.101)}
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\|g_1(u_h(x,t^n)) - g_1(u_h^n(x))\|_2 &= \left\|\frac{\nabla F_1(u_h(x,t^n))}{\mathcal{I}F_1(u_h(t^n))} - \frac{\nabla F_1(u_h^n(x))}{\mathcal{I}F_1(u_h^n)}\right\|_2 && \text{(I.102)}\\
&= \left\|\frac{\mathcal{I}F_1(u_h^n)\nabla F_1(u_h(x,t^n)) - \mathcal{I}F_1(u_h(t^n))\nabla F_1(u_h^n(x))}{\mathcal{I}F_1(u_h^n)\mathcal{I}F_1(u_h(t^n))}\right\|_2 && \text{(I.103)}\\
&\leqslant \frac{1}{c_1-\beta_{h,1}}\|\mathcal{I}F_1(u_h^n)\nabla F_1(u_h(x,t^n)) - \mathcal{I}F_1(u_h(t^n))\nabla F_1(u_h^n(x))\|_2 && \text{(I.104)}\\
&\leqslant \frac{1}{c_1-\beta_{h,1}}\|\mathcal{I}F_1(u_h(t^n))\left(\nabla F_1(u_h(x,t^n)) - \nabla F_1(u_h^n(x))\right) && \text{(I.105)}\\
&\qquad + \left(\mathcal{I}F_1(u_h(t^n)) - \mathcal{I}F_1(u_h^n)\right)\nabla F_1(u_h(x,t^n))\|_2 \\
&\leqslant \frac{1}{c_1-\beta_{h,1}}\Big[|\mathcal{I}F_1(u_h(t^n))|\,C_{df1}\,\|u_h(x,t^n) - u_h^n(x)\|_2 && \text{(I.106)}\\
&\qquad + C_{f1}\sqrt{\frac{|\Omega|}{c_1-\beta_{h,1}}}\|u_h(t^n) - u_h^n\|_{(L^2(\Omega))^p}\|\nabla F_1(u_h(x,t^n))\|_2\Big] \\
&\leqslant \frac{\|u_h(t^n) - u_h^n\|_{(L^2(\Omega))^p}}{c_1-\beta_{h,1}}\Big[C_{df1}\,|\mathcal{I}F_1(u_h(t^n))| && \text{(I.107)}\\
&\qquad + C_{f1}\sqrt{\frac{|\Omega|}{c_1-\beta_{h,1}}}\|\nabla F_1(u_h(x,t^n))\|_2\Big] \\
&\leqslant \frac{\|u_h(t^n) - u_h^n\|_{(L^2(\Omega))^p}}{c_1-\beta_{h,1}}\Big[C_{df1}\sup_{n\in[\![0,N]\!]}|\mathcal{I}F_1(u_h(t^n))| && \text{(I.108)}\\
&\qquad + C_{f1}\sqrt{\frac{|\Omega|}{c_1-\beta_{h,1}}}\sup_{p\in\mathcal{J}_h}\|\nabla F_1(p)\|_2\Big]
\end{aligned}
$$

So in the end

$$
\begin{aligned}
\|G_1(u_h(t^n)) - G_1(u_h^n)\|^2_{(L^2(\Omega))^p} &= \int_\Omega (G_1(u_h(t^n)) - G_1(u_h^n))\cdot(G_1(u_h(t^n)) - G_1(u_h^n)) && \text{(I.109)}\\
&= \int_\Omega \|g_1(u_h(x,t^n)) - g_1(u_h^n(x))\|_2^2\,dx && \text{(I.110)}\\
&\leqslant C_{g1}^2\|u_1 - u_2\|^2_{(L^2(\Omega))^p} && \text{(I.111)}
\end{aligned}
$$

with $C_{g1} = \dfrac{1}{c_1-\beta_{h,1}}\left(C_{df1}\sup_{n\in[\![0,N]\!]}|\mathcal{I}F_1(u_h(t^n))| + C_{f1}\sqrt{\dfrac{|\Omega|}{c_1-\beta_{h,1}}}\sup_{p\in\mathcal{J}_h}\|\nabla F_1(p)\|_2\right)$

which gives the expected result. A similar justification applies for $\bar{G}_2$.

3.3 Consistency

Let $\epsilon^n_{h,u} : \left(H^1(\Omega)\right)^p \longrightarrow \mathbb{R}$, $\epsilon^n_{h,z1} : H^1(\Omega) \longrightarrow \mathbb{R}$ and $\epsilon^n_{h,z2} : L^2(\Omega) \longrightarrow \mathbb{R}$ the truncation errors of the scheme such that

$$\begin{cases} \epsilon^n_{h,u}(u^*_h) = m\left(\delta^2 u_h(t^n), u^*_h\right) + a\left(\{u_h\}_\theta(t^n), u^*_h\right) \\ \qquad\qquad + \bar{G}_1\left(u_h(t^n), u^*_h, \mu^{1/2} z_{h,1}(t^n)\right) + \bar{G}_2\left(u_h(t^n), u^*_h, \mu^{1/2} z_{h,2}(t^n)\right) - f^n(u^*_h) & \text{(I.112a)} \\ \epsilon^n_{h,z1}(z^*_{h,1}) = \ell\left(\delta^{1/2} z_{h,1}(t^n), z^*_{h,1}\right) - \bar{G}_1\left(u_h(t^n), \delta\mu\, u_h(t^n), z^*_{h,1}\right) & \text{(I.112b)} \\ \epsilon^n_{h,z2}(z^*_{h,2}) = \ell\left(\delta^{1/2} z_{h,2}(t^n), z^*_{h,2}\right) - \bar{G}_2\left(u_h(t^n), \delta\mu\, u_h(t^n), z^*_{h,2}\right) & \text{(I.112c)} \end{cases}$$

Theorem 3.5 (Consistency)
If assumption 3.3 holds, the scheme I.32 is consistent in $O(\Delta t^2)$ and we have:

$$\forall\, (u^*_h, z^*_{h,1}, z^*_{h,2}) \in \mathcal{Q}_h \times \mathcal{Z}_{h,1} \times \mathcal{Z}_{h,2}, \quad \begin{cases} \left|\epsilon^n_{h,u}(u^*_h)\right| \leqslant C_{\epsilon,u}\Delta t^2 \left\|u^*_h\right\|_{(H^1(\Omega))^p} & \text{(I.113a)} \\ \left|\epsilon^n_{h,z1}(z^*_{h,1})\right| \leqslant C_{\epsilon,z1}\Delta t^2 \left\|z^*_{h,1}\right\|_{L^2(\Omega)} & \text{(I.113b)} \\ \left|\epsilon^n_{h,z2}(z^*_{h,2})\right| \leqslant C_{\epsilon,z2}\Delta t^2 \left\|z^*_{h,2}\right\|_{L^2(\Omega)} & \text{(I.113c)} \end{cases}$$

Corollary 3.1 (Modified Consistency)
If there is no type 2 nonlinear term in the equations and if assumption 3.3 holds, the scheme I.32 is consistent in $O(\Delta t^2)$ and we have:

$$\forall\, (u^*_h, z^*_{h,1}) \in \mathcal{Q}_h \times \mathcal{Z}_{h,1}, \quad \begin{cases} \left|\epsilon^n_{h,u}(u^*_h)\right| \leqslant C_{\epsilon,u}\Delta t^2 \left\|u^*_h\right\|_{(L^2(\Omega))^p} & \text{(I.114a)} \\ \left|\epsilon^n_{h,z1}(z^*_{h,1})\right| \leqslant C_{\epsilon,z1}\Delta t^2 \left\|z^*_{h,1}\right\|_{L^2(\Omega)} & \text{(I.114b)} \end{cases}$$

The difference is in the norm used for u^*_h: it is H^1 for the general case in the theorem above, and L^2 in this simplified case because there is no gradient nonlinear term.

Proof *Thanks to Taylor-Lagrange expansion we have $(\tau_i)_{1\leqslant i\leqslant 8} \in]t^{n-1}, t^{n+1}[^8$ so that:*

$$\begin{cases} \epsilon^n_{h,u}(u^*_h) = \dfrac{\Delta t^2}{12} m\left(u_h^{(4)}(\tau_1), u^*_h\right) + \theta\Delta t^2 a\left(u_h^{(2)}(\tau_2), u^*_h\right) \\ \qquad\qquad + \dfrac{\Delta t^2}{8}\bar{G}_1\left(u_h(\tau_3), u^*_h, z^{(2)}_{h,1}(\tau_4)\right) + \dfrac{\Delta t^2}{8}\bar{G}_2\left(u_h(\tau_3), u^*_h, z^{(2)}_{h,2}(\tau_5)\right) & \text{(I.115a)} \\ \epsilon^n_{h,z1}(z^*_{h,1}) = \dfrac{\Delta t^2}{24}\ell\left(z^{(3)}_{h,1}(\tau_6), z^*_{h,1}\right) - \dfrac{\Delta t^2}{3}\bar{G}_1\left(u_h(\tau_3), u_h^{(3)}(\tau_8), z^*_{h,1}\right) & \text{(I.115b)} \\ \epsilon^n_{h,z2}(z^*_{h,2}) = \dfrac{\Delta t^2}{24}\ell\left(z^{(3)}_{h,2}(\tau_7), z^*_{h,2}\right) - \dfrac{\Delta t^2}{3}\bar{G}_2\left(u_h(\tau_3), u_h^{(3)}(\tau_8), z^*_{h,2}\right) & \text{(I.115c)} \end{cases}$$

Then using continuity of $\bar{G}_i$ provided by proposition 3.3 and continuity of the bilinear forms we have:

$$\begin{cases} \left|\epsilon^n_{h,u}(u^*_h)\right| \leqslant \Delta t^2 \left[\frac{C_m}{12}\left\|u_h^{(4)}(\tau_1)\right\|_{(L^2(\Omega))^p}\|u^*_h\|_{(L^2(\Omega))^p} + \theta\left\|A_h u_h^{(2)}(\tau_2)\right\|_{(L^2(\Omega))^p}\|u^*_h\|_{(L^2(\Omega))^p}\right. \\ \qquad\qquad + \frac{C_{g1}}{8}\|u_h(\tau_3)\|_{(L^2(\Omega))^p}\left\|z^{(2)}_{h,1}(\tau_4)\right\|_{L^2(\Omega)}\|u^*_h\|_{(L^2(\Omega))^p} \qquad \text{(I.116a)} \\ \qquad\qquad \left. + \frac{C_{g2}}{8}\|u_h(\tau_3)\|_{(H^1(\Omega))^p}\left\|z^{(2)}_{h,2}(\tau_5)\right\|_{L^2(\Omega)}\|u^*_h\|_{(H^1(\Omega))^p}\right] \\ \left|\epsilon^n_{h,z1}(z^*_{h,1})\right| \leqslant \Delta t^2\left[\frac{C_\ell}{24}\left\|z^{(3)}_{h,1}(\tau_6)\right\|_{L^2(\Omega)} + \frac{C_{g1}}{3}\|u_h(\tau_3)\|_{(L^2(\Omega))^p}\left\|u_h^{(3)}(\tau_8)\right\|_{(L^2(\Omega))^p}\right]\left\|z^*_{h,1}\right\|_{L^2(\Omega)} \qquad \text{(I.116b)} \\ \left|\epsilon^n_{h,z2}(z^*_{h,2})\right| \leqslant \Delta t^2\left[\frac{C_\ell}{24}\left\|z^{(3)}_{h,2}(\tau_7)\right\|_{L^2(\Omega)} + \frac{C_{g2}}{3}\|u_h(\tau_3)\|_{(H^1(\Omega))^p}\left\|u_h^{(3)}(\tau_8)\right\|_{(H^1(\Omega))^p}\right]\left\|z^*_{h,2}\right\|_{L^2(\Omega)} \qquad \text{(I.116c)} \end{cases}$$

which proves the theorem with the constants:

$$\begin{cases} C_{\epsilon,u} = \frac{1}{12}C_m C_{0,4} + \theta C_{2,2} + \frac{1}{8}C_{g1}C_{0,1}C_{z1,2} + \frac{1}{8}C_{g2}C_{0,1}C_{z2,2} \qquad \text{(I.117a)} \\ C_{\epsilon,z1} = \frac{1}{24}C_\ell C_{z1,3} + \frac{1}{3}C_{g1}C_{0,1}C_{0,3} \qquad \text{(I.117b)} \\ C_{\epsilon,z2} = \frac{1}{24}C_\ell C_{z2,3} + \frac{1}{3}C_{g2}C_{0,1}C_{1,3} \qquad \text{(I.117c)} \end{cases}$$

with C_m and C_ℓ the continuity constants of m and ℓ and also the constants of assumption 3.3.

Remark 3.1 *It is actually possible to assess less space regularity on the semi-discrete solution by using a discrete summation by parts on the term $a(u_h^{(2)}, u_h^*)$ as done in [Chabassier and Imperiale, 2021]. $(p,q) = (1,3)$ in assumption 3.3 is sufficient and $(p,q) = (2,2)$ would not be required.*

3.4 Convergence for type 1 nonlinear terms

In this section we study the convergence of a scheme *without a nonlinear term of type 2*.

The simplified scheme is:
Seek $(u_h, z_{h,1}, z_{h,2}) \in \mathcal{Q}_h \times \mathcal{Z}_{h,1} \times \mathcal{Z}_{h,2}$ so that for all $(u_h^*, z_{h,1}^*, z_{h,2}^*) \in \mathcal{Q}_h \times \mathcal{Z}_{h,1} \times \mathcal{Z}_{h,2}$ and all $n \in [\![0, N]\!]$

$$\begin{cases} \left(\delta^2 u_h, u_h^*\right) + a\left(\{u_h\}_\theta^n, u_h^*\right) + \bar{G}_1\left(u_h^n, u_h^*, \mu^{1/2} z_{h,1}\right) = f^n(u_h^*) & \text{(I.118a)} \\ \ell\left(\delta^{1/2} z_{h,1}, z_{h,1}^*\right) = \bar{G}_1\left(u_h^n, \delta\mu\, u_h, z_{h,1}^*\right) & \text{(I.118b)} \end{cases}$$

Let $e_u^n = u_h(t^n) - u_h^n$ and $e_z^{n+1/2} = z_{h,1}(t^{n+1/2}) - z_{h,1}^{n+1/2}$ the time-discretization errors.

Theorem 3.7 (Convergence)
If the CFL condition 3.2 is satisfied and if the scheme is consistent in the sense of corollary 3.1, then the IEQ scheme I.118 converges with order 2 and:

$$\begin{cases} \|e_u^n\|_{(L^2(\Omega))^p} \leqslant 2\gamma\left(2\gamma C_{\epsilon,u} + C_{\epsilon,z1}\right) T^2 \Delta t^2\, e^{CT^2} & \text{(I.119a)} \\ \left\|e_z^{n+1/2}\right\|_{L^2(\Omega)} \leqslant \left(2\gamma C_{\epsilon,u} + C_{\epsilon,z1}\right) T \Delta t^2\, e^{CT^2} & \text{(I.119b)} \end{cases}$$

with

$$C = 2\gamma C_{g1}\left(2\gamma \max_{n\in[\![0,N]\!]} \left\|z_{h,1}^{n+1/2}\right\|_{L^2(\Omega)} + \max_{n\in[\![0,N]\!]} \|\delta\mu\, u_h\|_{(L^2(\Omega))^p}\right) \quad \text{(I.120)}$$

and $\gamma = \frac{C_P^{-1/2} + \frac{1}{2} C_K^{-1/2}}{c_m}$.

The result for SAV is the exact same with an absolute value $|\cdot|$ *instead of the* $\|\cdot\|_{L^2(\Omega)}$ *norm on the* z *quantities.*

Proof
We first subtract the scheme I.118 to the truncation errors I.112:

$$\begin{cases} m\left(\delta^2 e_u, u_h^*\right) + a\left(\{e_u\}_\theta^n, u_h^*\right) + \bar{G}_1\left(u_h(t^n), u_h^*, \mu^{1/2} z_{h,1}(t^n)\right) - \bar{G}_1\left(u_h^n, u_h^*, \mu^{1/2} z_{h,1}\right) = \epsilon_{h,u}^n(u_h^*) & \text{(I.121a)} \\ \ell\left(\delta^{1/2} e_{z1}, z_{h,1}^*\right) = \bar{G}_1\left(u_h(t^n), \delta\mu\, u_h(t^n), z_h^*\right) - \bar{G}_1\left(u_h^n, \delta\mu\, u_h, z_{h,1}^*\right) + \epsilon_{h,z1}^n(z_{h,1}^*) & \text{(I.121b)} \end{cases}$$

Using $u_h^* = \delta\mu\, e_u$ *and* $z_{h,1}^* = \mu^{1/2} e_{z1}$ *we obtain an energy identity with the errors:*

$$\begin{cases} \delta^{1/2}\mathcal{E}_e = \epsilon_{h,u}^n\left(\delta\mu\, e_u\right) + \epsilon_{h,z1}^n\left(\mu^{1/2} e_{z1}\right) + R_1^n + R_2^n & \text{(I.122a)} \\ \mathcal{E}_e^{n+1/2} = \frac{1}{2}\widetilde{m}\left(\delta e_u, \delta e_u\right) + \frac{1}{2} a\left(\mu e_u, \mu e_u\right) + \frac{1}{2}\ell\left(e_{z1}^{n+1/2}, e_{z1}^{n+1/2}\right) & \text{(I.122b)} \\ R_1^n = \bar{G}_1\left(u_h(t^n), \delta\mu\, e_u, \mu^{1/2} z_{h,1}\right) - \bar{G}_1\left(u_h^n, \delta\mu\, e_u, \mu^{1/2} z_{h,1}\right) & \text{(I.122c)} \\ R_2^n = -\bar{G}_1\left(u_h(t^n), \delta\mu\, u_h, \mu^{1/2} e_{z1}\right) + \bar{G}_1\left(u_h^n, \delta\mu\, u_h, \mu^{1/2} e_{z1}\right) & \text{(I.122d)} \end{cases}$$

Now using proposition 3.3 we can have:

$$\begin{cases} |R_1^n| \leqslant C_{g1}\, \|e_u^n\|_{(L^2(\Omega))^p}\, \|\delta\mu\, e_u\|_{(L^2(\Omega))^p}\, \left\|\mu^{1/2} z_{h,1}\right\|_{L^2(\Omega)} & \text{(I.123a)} \\ |R_2^n| \leqslant C_{g1}\, \|e_u^n\|_{(L^2(\Omega))^p}\, \|\delta\mu\, u_h\|_{(L^2(\Omega))^p}\, \left\|\mu^{1/2} e_{z1}\right\|_{L^2(\Omega)} & \text{(I.123b)} \end{cases}$$

and to estimate the error $\|e_u^n\|_{(L^2(\Omega))^p}$ *with the error-energy we write:*

$$\begin{aligned} \left\|\Pi_K e_u^{n+1}\right\|_m &\leqslant \|\Pi_K e_u^n\|_m + \Delta t \|\Pi_K \delta e_u\|_m && \text{(I.124)} \\ &\leqslant \|\Pi_K e_u^n\|_m + \Delta t C_K^{-1/2}\sqrt{2\mathcal{E}_e^{n+1/2}} && \text{(I.125)} \end{aligned}$$

and

$$\begin{aligned} \left\|\Pi_P e_u^{n+1}\right\|_m &\leqslant \|\Pi_P e_u^n\|_m + 2\|\Pi_P \mu e_u\|_m && \text{(I.126)} \\ &\leqslant \|\Pi_P e_u^n\|_m + 2\Delta t C_P^{-1/2}\sqrt{2\mathcal{E}_e^{n+1/2}} && \text{(I.127)} \end{aligned}$$

which implies with telescopic sums that

$$c_m \|e_u^n\|_{(L^2(\Omega))^p} \leqslant \|e_u^n\|_m \leqslant \|\Pi_K e_u^n\|_m + \|\Pi_P e_u^n\|_m \leqslant \sqrt{2}\left\|e_u^0\right\|_m + 2c_m\gamma\sqrt{2}\Delta t\sum_{i=0}^{n-1}\sqrt{\mathcal{E}_e^{i+1/2}} \tag{I.128}$$

and we also assess in the following that the initial error $e_u^0 = 0$.

Let's sum up the bounds we use:

$$\begin{cases} \left\|\mu^{1/2} z_{h,1}\right\|_{L^2(\Omega)} \leqslant \max\limits_{n\in[\![0,N]\!]} \left\|z_{h,1}^{n+1/2}\right\|_{L^2(\Omega)} := Z_{1,max} \quad \textit{because of I.62d} & \text{(I.129a)} \\ \left\|\mu^{1/2} e_{z1}\right\|_{L^2(\Omega)} \leqslant \frac{1}{2}\sqrt{2}\left(\sqrt{\mathcal{E}_e^{n+1/2}} + \sqrt{\mathcal{E}_e^{n-1/2}}\right) & \text{(I.129b)} \\ \|\delta\mu\, u_h\|_{(L^2(\Omega))^p} \leqslant \max\limits_{n\in[\![0,N]\!]} \|\delta\mu\, u_h\|_{(L^2(\Omega))^p} := \dot{U}_{max} \quad \textit{because of I.62b} & \text{(I.129c)} \\ \|\delta\mu\, e_u\|_{(L^2(\Omega))^p} \leqslant \gamma\sqrt{2}\left(\sqrt{\mathcal{E}_e^{n+1/2}} + \sqrt{\mathcal{E}_e^{n-1/2}}\right) & \text{(I.129d)} \\ |R_1^n| \leqslant 4\gamma^2 C_{g1}\Delta t\, Z_{1,max}\left(\sqrt{\mathcal{E}_e^{n+1/2}} + \sqrt{\mathcal{E}_e^{n-1/2}}\right)\sum\limits_{j=0}^{n-1}\sqrt{\mathcal{E}_e^{j+1/2}} & \text{(I.129e)} \\ |R_2^n| \leqslant 2\gamma C_{g1}\Delta t\, \dot{U}_{max}\left(\sqrt{\mathcal{E}_e^{n+1/2}} + \sqrt{\mathcal{E}_e^{n-1/2}}\right)\sum\limits_{j=0}^{n-1}\sqrt{\mathcal{E}_e^{j+1/2}} & \text{(I.129f)} \end{cases}$$

We now apply those bounds to I.122a we have:

$$\begin{aligned} &\frac{1}{\Delta t}\left(\mathcal{E}_e^{n+1/2} - \mathcal{E}_e^{n-1/2}\right) \leqslant \gamma\sqrt{2}C_{\epsilon,u}\Delta t^2\left(\sqrt{\mathcal{E}_e^{n+1/2}} + \sqrt{\mathcal{E}_e^{n-1/2}}\right) + \frac{1}{2}\sqrt{2}C_{\epsilon,z1}\Delta t^2\left(\sqrt{\mathcal{E}_e^{n+1/2}} + \sqrt{\mathcal{E}_e^{n-1/2}}\right) \\ &+ \Delta t\underbrace{\left[4\gamma^2 C_{g1}Z_{1,max} + 2\gamma C_{g1}\dot{U}_{max}\right]}_{C}\left(\sqrt{\mathcal{E}_e^{n+1/2}} + \sqrt{\mathcal{E}_e^{n-1/2}}\right)\sum_{j=0}^{n-1}\sqrt{\mathcal{E}_e^{j+1/2}} \end{aligned} \tag{I.130}$$

We can now simplify with $\sqrt{\mathcal{E}_e^{n+1/2}} + \sqrt{\mathcal{E}_e^{n-1/2}}$ *and sum from 1 to n:*

$$\sqrt{\mathcal{E}_e^{n+1/2}} \leqslant \underbrace{\gamma\sqrt{2}\Delta t^2 C_{\epsilon,u}T + \frac{\Delta t^2}{\sqrt{2}}C_{\epsilon,z1}T}_{=A\Delta t^2} + C\Delta t^2\sum_{i=0}^{n-1}\sum_{j=0}^{i}\sqrt{\mathcal{E}_e^{j+1/2}} \tag{I.131}$$

We can now apply the modified Gronwall lemma A.1.

$$\sqrt{\mathcal{E}_e^{n+1/2}} \leqslant A\Delta t^2\, e^{C\Delta t^2 n_0(n_0+1)} \leqslant A\Delta t^2\, e^{CT^2} \tag{I.132}$$

and use this result in I.128 which gives:

$$c_m \left\|e_u^n\right\|_{(L^2(\Omega))^p} \leqslant \left\|e_u^n\right\|_m \leqslant 2\gamma\sqrt{2}TA\Delta t^2\, e^{CT^2} \tag{I.133}$$

and then finally

$$\left\|e_u^n\right\|_{(L^2(\Omega))^p} \leqslant 2\gamma\left(2\gamma C_{\epsilon,u} + C_{\epsilon,z1}\right)T^2\Delta t^2\, e^{CT^2} \tag{I.134}$$

3.5 Convergence for type 2 nonlinear terms

To prove convergence for a scheme with a nonlinear term of type 2 like $F_2(\nabla u)$, we would proceed just like before and write an energy on the errors:

$$\begin{cases} \delta^{1/2}\mathcal{E}_e = \epsilon_{h,u}^n\left(\delta\mu\, e_u\right) + \epsilon_{h,z2}^n\left(\mu^{1/2}e_{z2}\right) + R_1^n + R_2^n & \text{(I.135a)} \\ \mathcal{E}_e^{n+1/2} = \dfrac{1}{2}\widetilde{m}\left(\delta e_u, \delta e_u\right) + \dfrac{1}{2}a\left(\mu e_u, \mu e_u\right) + \dfrac{1}{2}\ell\left(e_{z2}^{n+1/2}, e_{z2}^{n+1/2}\right) & \text{(I.135b)} \\ R_1^n = \bar{G}_2\left(u_h(t^n), \delta\mu\, e_u, \mu^{1/2}z_{h,2}\right) - \bar{G}_2\left(u_h^n, \delta\mu\, e_u, \mu^{1/2}z_{h,2}\right) & \text{(I.135c)} \\ R_2^n = -\bar{G}_2\left(u_h(t^n), \delta\mu\, u_h, \mu^{1/2}e_{z2}\right) + \bar{G}_2\left(u_h^n, \delta\mu\, u_h, \mu^{1/2}e_{z2}\right) & \text{(I.135d)} \end{cases}$$

Unfortunately, proposition 3.4 gives

$$\begin{cases} \left|R_1^n\right| \leqslant C_{g2}\left\|e_u^n\right\|_{(H^1(\Omega))^p}\left\|\delta\mu\, e_u\right\|_{(H^1(\Omega))^p}\left\|\mu^{1/2}z_{h,2}\right\|_{L^2(\Omega)} & \text{(I.136a)} \\ \left|R_2^n\right| \leqslant C_{g2}\left\|e_u^n\right\|_{(H^1(\Omega))^p}\left\|\delta\mu\, u_h\right\|_{(H^1(\Omega))^p}\left\|\mu^{1/2}e_{z2}\right\|_{L^2(\Omega)} & \text{(I.136b)} \end{cases}$$

with H^1 norms which we are not able to control by the energy.
We could try three solutions to bound the errors:

- $\left\|e_u^n\right\|_{H^1} \leqslant \left\|\mu e_u\right\|_{H^1} + \frac{\Delta t}{2}\left\|\delta e_u\right\|_{H^1}$ but the norm $\left\|\delta e_u\right\|_{H^1}$ does not appear in the energy.
- $\left\|e_u^n\right\|_{H^1} \leqslant \left\|e_u^{n-1}\right\|_{H^1} + \Delta t\left\|\delta e_u^{n-1/2}\right\|_{H^1} \leqslant \left\|e_u^0\right\|_{H^1} + \Delta t\sum\left\|\delta e_u^{j-1/2}\right\|_{H^1}$ but the norms $\left\|\delta e_u^{j-1/2}\right\|_{H^1}$ do not appear in the energy.
- $\left\|e_u^n\right\|_{H^1} \leqslant \left\|e_q^{n-1}\right\|_{H^1} + 2\left\|\mu e_u^{n-1/2}\right\|_{H^1} \leqslant \left\|e_u^0\right\|_{H^1} + 2\sum\left\|\mu e_u^{j-1/2}\right\|_{H^1}$ which can be controlled by the energy but the lack of Δt in front of the sum makes it inconsistant with a continuous intergral and leads to a less-than-quadratic result.
- Also the decomposition with the projectors Π_K and Π_P is not possible with H^1 norms to the best of our knowledge.

An inverse inequality assessing equivalence of norms in finite dimension $\left\|.\right\|_{H^1} \leqslant c_h\left\|.\right\|_{L^2}$ is sufficient to complete a time convergence proof following the previous proof sketch. However it does not give proper space-time convergence results since the constant c_h diverges for $h \to 0$.

The numerical convergence results in the next part about the piano string seem to show that such a scheme with type 2 nonlinear term does not space-time converge without extra assumptions.

4 Phase formulation of quadratized schemes

As a reminder, the abstract problem I.1 is

$$\partial_t^2 M u - \mathrm{Div}\left(A\nabla u + \nabla F_2(\nabla u)\right) + \nabla F_1(u) = f \tag{I.137}$$

It can be reformulated as a phase formulation with an extra field p such that

$$\begin{cases} p = \partial_t u & \text{(I.138a)} \\ M\partial_t p - \mathrm{Div}\left(A\nabla u + \nabla F_2(\nabla u)\right) + \nabla F_1(u) = f & \text{(I.138b)} \end{cases}$$

The quadratization techniques still apply and can derive modified P-IEQ and P-SAV schemes based on I.138.

4.1 Phase P-IEQ numerical scheme

Weak semi-discrete formulation of the quadratized equations are given directly since the quadratization process does not change at all.
p is sought in $\left(H^1(\Omega)\right)^p$ and p_h in $\mathcal{Q}_h$ similarly to the fields u and u_h.

Find $(u_h, p_h) \in \mathcal{Q}_h^2$ and $z_{h,i} \in \mathcal{Z}_{h,i}$ such that for all $(u_h^*, p_h^*) \in \mathcal{Q}_h^2$ and all $z_{h,i}^* \in \mathcal{Z}_{h,i}$:

$$\begin{cases} \ell(p_h, u_h^*) = \ell(\partial_t u_h, u_h^*) & \text{(I.139a)} \\ m(\partial_t p_h, p_h^*) + a(u_h, p_h^*) + \bar{G}_1(u_h, p_h^*, z_{h,1}) + \bar{G}_2(u_h, p_h^*, z_{h,2}) = f(p_h^*) & \text{(I.139b)} \\ \ell(\partial_t z_{h,1}, z_{h,1}^*) = \bar{G}_1(u_h, p_h, z_{h,1}^*) & \text{(I.139c)} \\ \ell(\partial_t z_{h,2}, z_{h,2}^*) = \bar{G}_2(u_h, p_h, z_{h,2}^*) & \text{(I.139d)} \end{cases}$$

Which leads to the fully-discrete scheme inspired from [Jiang et al., 2019] where all the variables P_h, U_h and $Z_{h,i}$ are discretized on the same time grid t^n

Numerical Scheme 4.1 (P-IEQ Time Scheme)

$$\begin{cases} \mu P_h = \delta U_h & \text{(I.140a)} \\ M_h \delta P_h + K_h \mu U_h + {}^t\mathbb{G}_1(\pi U_h)\mu Z_{h,1} + {}^t\mathbb{G}_2(\pi U_h)\mu Z_{h,2} = F_h^{n+1/2} & \text{(I.140b)} \\ L_{h,1}\delta Z_{h,1} = \mathbb{G}_1(\pi U_h)\mu P_h & \text{(I.140c)} \\ L_{h,2}\delta Z_{h,2} = \mathbb{G}_2(\pi U_h)\mu P_h & \text{(I.140d)} \end{cases}$$

where πU_h is any second order consistent extrapolation of $U_h(t^{n+1/2})$. For example $\pi U_h = \frac{1}{2}\left(3U_h^n - U_h^{n-1}\right)$.

Note that this scheme has the same computational complexity as I.27 as soon as πU_h is explicit.

This I.140 P-IEQ scheme has the discrete energy conservation law

Theorem 4.1 (Discrete energy Identity of P-IEQ scheme)

$$\delta\mathcal{E}_h = F_h^{n+1/2} \cdot \mu P_h \tag{I.141}$$

with

$$\mathcal{E}_h^n = \frac{1}{2} M_h P_h^n \cdot P_h^n + \frac{1}{2} K_h U_h^n \cdot U_h^n + \frac{1}{2} L_{h,1} Z_{h,1}^n \cdot Z_{h,1}^n + \frac{1}{2} L_{h,2} Z_{h,2}^n \cdot Z_{h,2}^n \tag{I.142}$$

4.2 Phase P-SAV numerical scheme

Find $(u_h, p_h) \in \mathcal{Q}_h^2$ and $(z_{h,1}, z_{h,2}) \in \mathbb{R}^2$ such that for all $(u_h^*, p_h^*) \in \mathcal{Q}_h^2$:

$$\begin{cases} \ell(p_h, u_h^*) = \ell(\partial_t u_h, u_h^*) & \text{(I.143a)} \\ m(\partial_t p_h, p_h^*) + a(u_h, p_h^*) + z_{h,2}\bar{G}_2(u_h, p_h^*) + z_{h,1}\bar{G}_1(u_h, p_h^*) = f(p_h^*) & \text{(I.143b)} \\ \dot{z}_{h,1} = \bar{G}_1(u_h, p_h) & \text{(I.143c)} \\ \dot{z}_{h,2} = \bar{G}_2(u_h, p_h) & \text{(I.143d)} \end{cases}$$

Which leads to the fully-discrete scheme inspired from [Jiang et al., 2019]

Numerical Scheme 4.2 (P-SAV Time Scheme)

$$\begin{cases} \mu P_h = \delta U_h & \text{(I.144a)} \\ M_h \delta P_h + K_h \mu U_h + \mu z_{h,1}\, \mathbb{G}_1(\pi U_h) + \mu z_{h,2}\, \mathbb{G}_2(\pi U_h) = F_h^{n+1/2} & \text{(I.144b)} \\ \delta z_{h,1} = \mathbb{G}_1(\pi U_h) \cdot \mu P_h & \text{(I.144c)} \\ \delta z_{h,2} = \mathbb{G}_2(\pi U_h) \cdot \mu P_h & \text{(I.144d)} \end{cases}$$

where πU_h is any second order consistent extrapolation of $U_h(t^{n+1/2})$. For example $\pi U_h = \frac{1}{2}\left(3U_h^n - U_h^{n-1}\right)$.

Note that this scheme has the same computational complexity as I.44 as soon as πU_h is explicit.

This I.144 P-SAV scheme has the discrete energy conservation law

Theorem 4.2 (Discrete energy Identity of P-SAV scheme)

$$\delta \mathcal{E}_h = F_h^{n+1/2} \cdot \mu P_h \tag{I.145}$$

with

$$\mathcal{E}_h^n = \frac{1}{2} M_h P_h^n \cdot P_h^n + \frac{1}{2} K_h U_h^n \cdot U_h^n + \frac{1}{2}\left(z_{h,1}^n\right)^2 + \frac{1}{2}\left(z_{h,2}^n\right)^2 \tag{I.146}$$

Remark 4.1 *With these schemes it is possible to derive H^1 bound for the time discretization error e_u for nonlinear terms of type 1 [Jiang et al., 2019].*

Part II
Simulation of the nonlinear piano string

The string is the key element of the piano which is the source of all the vibrations. After those vibrations are created they are transmitted to a soundboard and are radiated in the air.
In the perspective of creating artificial sounds with computer simulation and because of this particular role, precise and fast numerical methods are required for solving the string equations.
[Chabassier and Joly, 2010], [Chabassier, 2012] and [Ducceschi and Bilbao, 2022][Ducceschi et al., 2022] use two different approaches. One with Discrete Gradient scheme and quasi-Newton solver which can be really slow and have trouble converging for highly nonlinear cases, and the other with quadratized scheme which are faster but for which no real mathematical studies were carried out.
In the following we apply the theoretical results of I concerning SAV to the piano string and relate them to numerical simulations. We also give some comparisons between the Discerte Gradient (GRAD) and the SAV schemes.

1 Piano string model

A non-stiff non-damped piano string of length L can be modeled with the general equation given in [Chabassier, 2012]:
For $x \in [0, L]$ and $t \in [0, T]$, $q \equiv q(x, t) \in \mathbb{R}^2$ verifies

$$M\partial_t^2 q - \partial_x \left(A\partial_x q + \nabla\mathcal{U}(\partial_x q)\right) = f(x, t) \tag{II.1}$$

For the so-called Geometrically Exact Model (GEM) we have:

$$q = \begin{pmatrix} u \\ v \end{pmatrix}, \quad M = \begin{pmatrix} \rho S & 0 \\ 0 & \rho S \end{pmatrix}, \quad A = \begin{pmatrix} ES & 0 \\ 0 & ES \end{pmatrix}, \quad \mathcal{U}(u, v) = (ES - T_0)\left[(1 + v) - \sqrt{u^2 + (1 + v)^2}\right] \tag{II.2}$$

with the density ρ, the section S, the Young modulus E and the tension at rest T_0.
u and v stand for the transverse and longitudinal displacement of the string.

Note that adding an affine contribution to $\mathcal{U}$ does not change the equation II.1.

At $t = 0$ we consider the string at rest:

$$\forall x \in [0, L], \begin{cases} q(x, t = 0) = 0 \\ \partial_t q(x, t = 0) = 0 \end{cases} \tag{II.3}$$

and we set Dirichlet boundary conditions:

$$\forall t \in [0, T],\ q(0, t) = q(L, t) = 0 \tag{II.4}$$

Theorem 1.1 (Classical Solution)
The Geometrically Exact Model II.1 with non-zero initial conditions with no source term and boundary conditions II.4 has a unique solution $q \in \mathscr{C}^2(\mathbb{R} \times \mathbb{R}^+)$.

Proof *The proof is given in [Chabassier, 2012] using the theorem of [Ta-Tsien et al., 1994] for small enough initial conditions and no source term.*

Proposition 1.1 (A-priori estimates)
Let q be a solution of II.1 with conditions II.3 and II.4 and let $f \in L^1\left([0,T], L^2\right)$.
For all $t \in [0,T]$ we have

$$\|\dot{q}(\cdot,t)\|_{L^2} \leqslant \frac{1}{\rho S}\int_0^t \|f(\cdot,s)\|_{L^2}\, ds \tag{II.5}$$

$$|q(\cdot,t)|_{H^1} \leqslant \frac{1}{\sqrt{\rho S T_0}}\int_0^t \|f(\cdot,s)\|_{L^2}\, ds \tag{II.6}$$

$$\|q(\cdot,t)\|_{L^2} \leqslant \frac{1}{\rho S}\int_0^t (t-s)\,\|f(\cdot,s)\|_{L^2}\, ds \tag{II.7}$$

Proof *See [Chabassier, 2012] for details about the proof.*

2 Properties of the nonlinear function

As mentioned in part I in remark 1.2, we can introduce a matrix $\alpha = \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix}$ such that II.1 writes

$$M\partial_t^2 q - \partial_x\left(\alpha A \partial_x q + \nabla \mathcal{U}_\alpha(\partial_x q)\right) = f(x,t) \tag{II.8}$$

with

$$\forall p \in \mathbb{R}^2, \quad \mathcal{U}_\alpha(p) = \frac{1}{2}(I_2 - \alpha)Ap\cdot p + \mathcal{U}(p) \tag{II.9}$$

Proposition 2.1 (Special value of α)
$\alpha^* = diag\left(\frac{T_0}{ES}, 1\right)$ is a particular value of α which makes the quadratic terms of $\mathcal{U}_\alpha$ vanish around $(0,0)$.

Proof

$$\begin{aligned}
\mathcal{U}_{\alpha^*}(u,v) &\sim \frac{1}{2}\begin{pmatrix} ES - T_0 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}\cdot\begin{pmatrix} u \\ v \end{pmatrix} && \text{(II.10)}\\
&\qquad + (T_0 - ES)\left[1 + \frac{1}{2}(2v + v^2 + u^2) - \frac{1}{8}(2v + v^2 + u^2)^2 - (1+v)\right] \\
&\sim \frac{1}{2}(ES - T_0)u^2 + (T_0 - ES)\left[\frac{1}{2}(u^2+v^2) - \frac{1}{2}v^2 - \frac{1}{8}(v^4 + u^4 + 4v^3 + 4vu^2 + 2v^2u^2)\right] && \text{(II.11)}\\
&\sim \frac{1}{8}(ES - T_0)\left[v^4 + u^4 + 4v^3 + 4vu^2 + 2v^2u^2\right] && \text{(II.12)}
\end{aligned}$$

Proposition 2.2 *The function $\mathcal{U}_\alpha$ verifies*

$$\forall p \in \mathbb{R}^2, \quad \|\nabla \mathcal{U}_\alpha(p)\|_2 \leqslant M_\alpha\left(1 + \|p\|_2\right) \tag{II.13}$$

with $M_\alpha = ES\max\left(1 - \max(\alpha), \frac{2T_0}{ES}\right)$

Proof $\nabla \mathcal{U}_\alpha(u,v) = (I_N - \alpha)A\begin{pmatrix} u \\ v \end{pmatrix} - (ES - T_0)\left[\frac{1}{\sqrt{u^2 + (1+v)^2}}\begin{pmatrix} u \\ 1+v \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right]$

Which gives that
$\|\nabla \mathcal{U}_\alpha(p)\|_2 \leqslant ES(1 - \alpha_{max})\,\|p\|_2 + 2|ES - T_0| \leqslant M_\alpha(1 + \|p\|_2)$

Proposition 2.3 *The nonlinear function $\mathcal{U}_\alpha$ and its gradient $\nabla\mathcal{U}_\alpha$ verify the Lipschitz assumptions 2.1.*

Proof *$\mathcal{U}_\alpha$ and $\nabla\mathcal{U}_\alpha$ are not well-defined at point $(0,-1)$ but are $\mathscr{C}^\infty$ everywhere else. So they are Lipschitzian on the bounded domain $\mathcal{I}_\nabla = \left\{(x,y)\in\mathbb{R}^2 \mid x^2+y^2<1\right\}$*

3 Numerical schemes for the piano string

The results of part I apply to the piano string with $F_1 = 0$ and $F_2 = \mathcal{U}_\alpha$.

3.1 Invariant Energy Quadratization (IEQ)

The auxiliary variable is defined as

$$z(x,t) = \sqrt{2\mathcal{U}_\alpha(\partial_x q(x,t)) + c} \tag{II.14}$$

and the auxiliary function

$$\forall p\in\mathbb{R}^2, \quad g_\alpha(p) = \frac{1}{\sqrt{2\mathcal{U}_\alpha(p)+c}}\nabla\mathcal{U}_\alpha(p) \tag{II.15}$$

The IEQ weak formulation of the piano string writes:
Seek $q\in\left(H_0^1([0,L])\right)^2$ and $z\in L^2([0,L])$ such that for all $q^*\in\left(H_0^1([0,L])\right)^2$ and all $z^*\in L^2([0,L])$:

$$\begin{cases} \displaystyle\int_0^L \partial_t^2 Mq\cdot q^* + \int_0^L A\partial_x q\cdot\partial_x q^* + \int_0^L zG_\alpha(\partial_x q)\cdot\partial_x q^* = \int_0^L f\cdot q^* & \text{(II.16a)} \\ \displaystyle\int_0^L \partial_t z z^* = \int_0^L z^* G_\alpha(\partial_x q)\cdot\partial_t\partial_x q & \text{(II.16b)} \end{cases}$$

Proposition 3.1 (Special value of α)
$\alpha^* = diag\left(\frac{T_0}{ES},1\right)$ is the only value of α for which $\nabla g_{\alpha^*}(0,0) = \begin{pmatrix}0\\0\end{pmatrix}$.

Proof

$$\nabla g_\alpha = \frac{\nabla^2\mathcal{U}_\alpha}{\sqrt{2\mathcal{U}_\alpha+c}} - \frac{\nabla\mathcal{U}_\alpha\ {}^t\nabla\mathcal{U}_\alpha}{\sqrt{2\mathcal{U}_\alpha+c}^3} \tag{II.17}$$

Elementary algebraic calculations show that

$$\nabla g_\alpha(0,0) = \frac{\nabla^2\mathcal{U}_\alpha(0,0)}{\sqrt{c}} = \frac{1}{\sqrt{c}}\left[(I_2-\alpha)A - (ES-T_0)\begin{pmatrix}1&0\\0&0\end{pmatrix}\right] \tag{II.18}$$

which concludes the proof.

From proposition 3.3 we can give expressions of the Lipschitz constant of g_α named $C_{g\alpha}$.

$$C_{g\alpha} = \frac{1}{c-\beta_h}\left[C_{\nabla\mathcal{U}_\alpha}\sup_{p\in\mathcal{J}_{\nabla,h}}\sqrt{2\mathcal{U}_\alpha(p)+c} + \frac{C_{\mathcal{U}_\alpha}}{\sqrt{c-\beta_h}}\sup_{p\in\mathcal{J}_{\nabla,h}}\|\nabla\mathcal{U}_\alpha(p)\|_2\right] \tag{II.19}$$

$\beta_h = -2\inf_{p\in\mathcal{I}_{\nabla,h}}\mathcal{U}_\alpha(p)$ is the minimal value of the nonlinear function and c is chosen so that $c-\beta_h>0$.
It also depends on the Lipschitz constants $C_{\mathcal{U}_\alpha}$ and $C_{\nabla\mathcal{U}_\alpha}$ of the nonlinear functions $\mathcal{U}_\alpha$ and $\nabla\mathcal{U}_\alpha$ defined by

$$\begin{cases} C_{\mathcal{U}_\alpha} = \sup_{p\in\mathcal{I}_{\nabla,h}}\|\nabla\mathcal{U}_\alpha(p)\|_2 & \text{(II.20a)} \\ C_{\nabla\mathcal{U}_\alpha} = \sup_{p\in\mathcal{I}_{\nabla,h}}\left|\left|\left|\nabla^2\mathcal{U}_\alpha(p)\right|\right|\right|_2 = \sup_{p\in\mathcal{I}_{\nabla,h}}\rho\left(\nabla^2\mathcal{U}_\alpha(p)\right) & \text{(II.20b)} \end{cases}$$

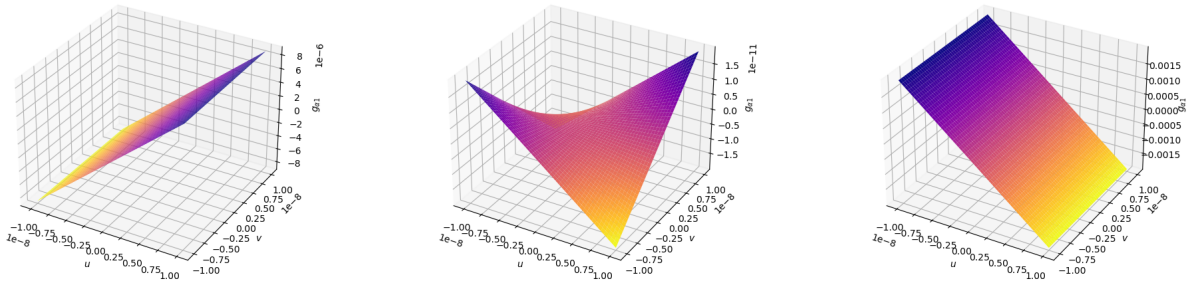

Figure 1: First component of function g_α for $\alpha_1 = 0$ (left), $\alpha_1 = \alpha_1^* = \frac{T_0}{ES}$ (center) and $\alpha_1 = 1$ (right).

in accordance with assumption 2.1 and with $\rho(\cdot)$ being the spectral radius.

The last proposition 3.1 shows that the nonlinear function g_{α^*} is flat around the origin $(0, 0)$ which means that the Lipschitz constant $C_{g\alpha}$ will tend to be smaller for this particular value α^*.

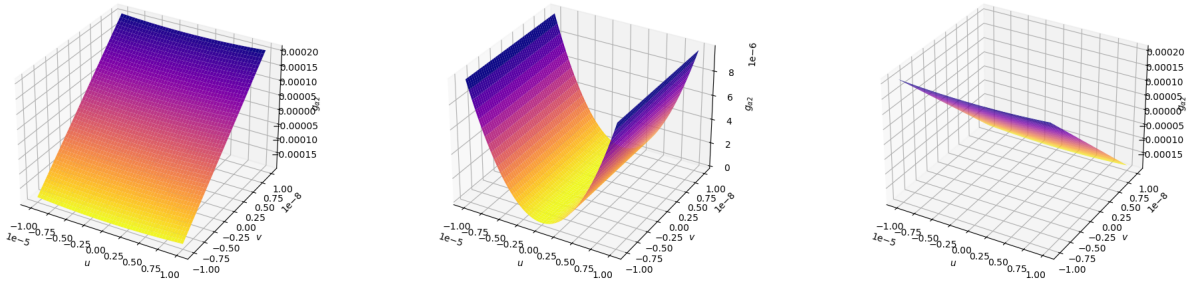

Figure 2: Second component of function g_α for $\alpha_2 = 0.9$ (left), $\alpha_2 = \alpha_2^* = 1$ (center) and $\alpha_2 = 1.1$ (right).

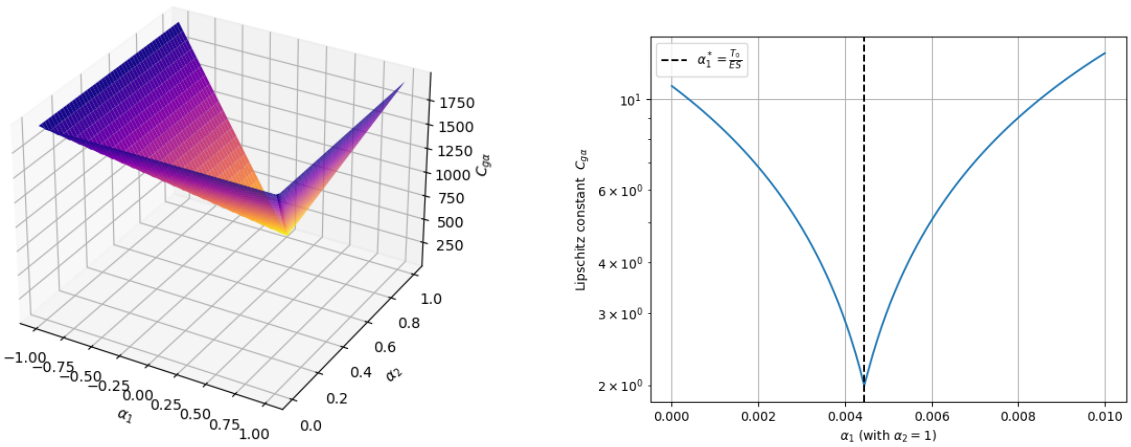

Figure 3: Values on the Lipschitz constant $C_{g\alpha}$ with respect to the value of α.

The smallest possible value for $C_{g\alpha}$ is indeed obtained with $\alpha = \text{diag}(\alpha_1, \alpha_2) = \text{diag}\left(\frac{T_0}{ES}, 1\right) = \alpha^*$ as shown on figure 3.

The auxiliary constant c also has an influence on the value of the Lipschitz constant. For $c = 0$, g_α is not even continuous around $(0, 0)$. The larger c gets, the smother g_α becomes.

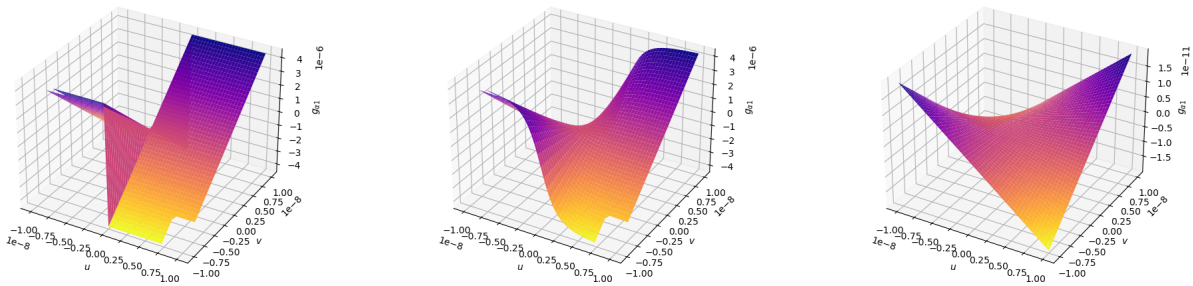

Figure 4: First component of function g_α for $c = 0$ (left), $c = 10^{-12}$ (center) and $c = 1$ (right).

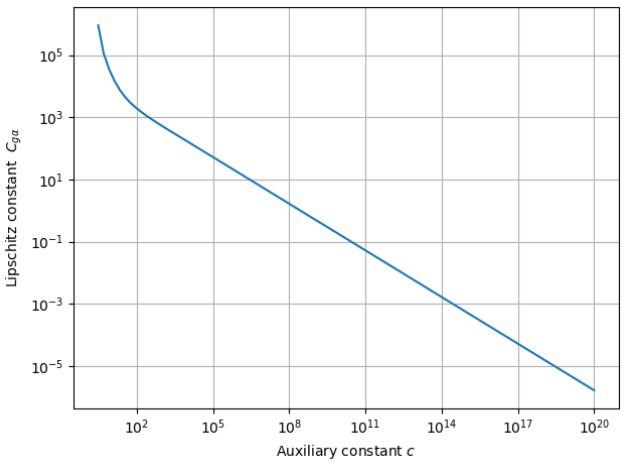

Figure 5: Values on the Lipshitz constant $C_{g\alpha}$ with respect to the value of c.

Since the string vibrations tend to be small, using α^* and a large auxiliary constant c in the simulations should make the convergence bounds more sharp in the theorem 3.7 and increase precision of the scheme.

Note that the figures of this section shows Lipschitz constants computed on a rectangular set centered on (0,0) $\mathcal{I}_{\nabla,h} = \mathcal{J}_{\nabla,h}$. In practice, the solutions live in a more restricted domain so the Lipschitz constant may vary from one simulation to another.

3.2 Scalar Auxiliary Variable (SAV)

The auxiliary variable is defined as

$$z(t) = \sqrt{2\int_0^L \mathcal{U}_\alpha(\partial_x q(x,t))dx + c} \tag{II.21}$$

and the auxiliary functions

$$\forall q \in \left(H_0^1([0,L])\right)^2, \forall (x,t) \in [0,L]\times[0,T], \quad g_\alpha(q(x,t)) = \frac{1}{\sqrt{2\int_\Omega \mathcal{U}_\alpha(\partial_x q(s,t))ds + c}} \nabla \mathcal{U}_\alpha(\partial_x q(x,t)) \tag{II.22}$$

$$\forall (q,q^*) \in \left(H_0^1([0,L])\right)^2 \times \left(H_0^1([0,L])\right)^2, \quad \bar{G}_\alpha\left(\partial_x q, \partial_x q^*\right) = \int_0^L g_\alpha(q(x,t)) \cdot \partial_x q^*(x)dx \tag{II.23}$$

The SAV weak formulation of the piano string writes:
Seek $q \in \left(H_0^1([0,L])\right)^2$ and $z : [0,T] \longrightarrow \mathbb{R}$ such that for all $q^* \in \left(H_0^1([0,L])\right)^2$:

$$\begin{cases} \displaystyle\int_0^L \partial_t^2 M q \cdot q^* + \int_0^L A\partial_x q \cdot \partial_x q^* + z\bar{G}_\alpha(\partial_x q, \partial_x q^*) = \int_0^L f \cdot q^* & \text{(II.24a)} \\ \partial_t z = \bar{G}_\alpha(\partial_x q, \partial_t \partial_x q) & \text{(II.24b)} \end{cases}$$

The SAV Lipschitz constant is a lot more complex to compute because it depends on the semi-discrete solution. We do not show the plots of the previous section with SAV.

4 Numerical results

For the following simulations we use a piano string whose physical parameters are

L (m)	S (m^2)	ρ (kg/m^3)	T_0 (N)	E (Pa)
1.0	$9.7993 \cdot 10^{-7}$	7850	880	$2.02 \cdot 10^{11}$

It gives a 169 Hz fundamental transverse frequency (which is approximately the note E_3) and 2536 Hz for the longitudinal frequency.

We use a $\mathscr{C}^\infty$ source in space and time applied on the transverse direction and defined by

$$f(x,t) = \begin{cases} A\, e^{1-\frac{1}{1-\left(\frac{x-x_0}{\sigma_x}\right)^2}}\, e^{1-\frac{1}{1-\left(\frac{t-t_0}{\sigma_t}\right)^2}} & \text{if} \quad (x,t) \in [x_0-\sigma_x, x_0+\sigma_x] \times [t_0-\sigma_t, t_0+\sigma_t] \\ 0 & \text{elsewhere} \end{cases}$$

with $A = 1000$, $t_0 = 0.3$ms, $\sigma_t = 0.2$ms, $x_0 = L/4$ and $\sigma_x = L/10$, meaning that the source starts at $t = 0.1$ms and ends at $t = 0.5$ms.

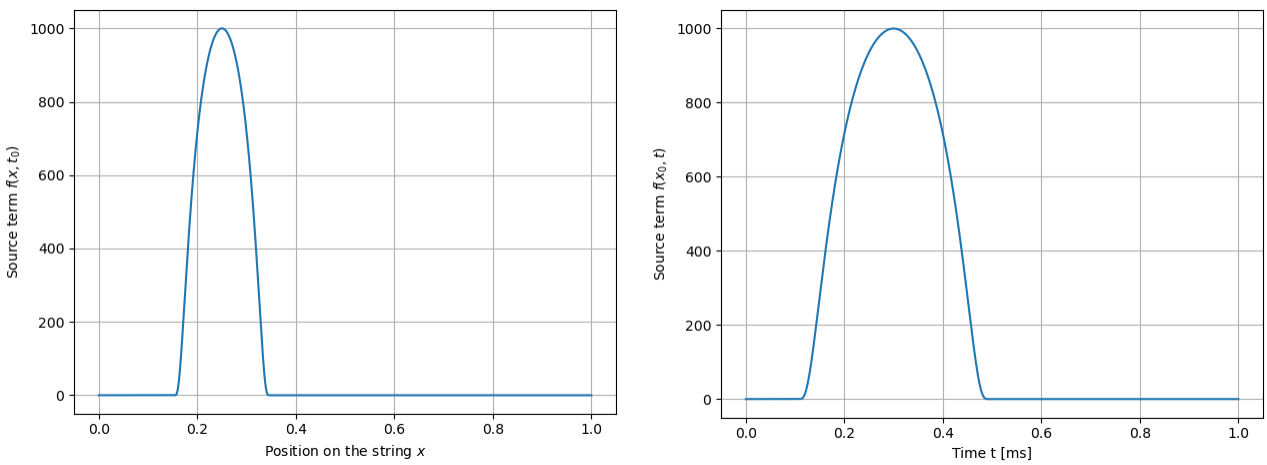

Figure 6: Source function for fixed time (left) and fixed space (right).

Until contrary mention a $\theta = 1/4$ scheme is used which is unconditionally stable, along with fourth order finite elements and the $\alpha = \alpha^*$ stabilization.

The schemes presented in part I have been implemented in the C++ Finite Elements solver MONTJOIE[1].

Along with the IEQ and SAV schemes we also present some numerical results obtained with a discrete gradient scheme (GRAD) solved with quasi-Newton iterations whose details can be found in [Chabassier, 2012].

The reference solution mentioned in this section is computed with the discrete gradient scheme (GRAD) with 2048 elements of order 4 with time step $\Delta t = 2.10^{-9}$s and with `'long double'` precision using MFPR[2] multiple precision library.

[1] `https://www.math.u-bordeaux.fr/~durufle/montjoie/index.php`
[2] `https://www.mpfr.org`

4.1 Solutions of the schemes

Figure 7 shows converged SAV simulations of the considered piano string on a 20ms interval at point $x = L/4$.

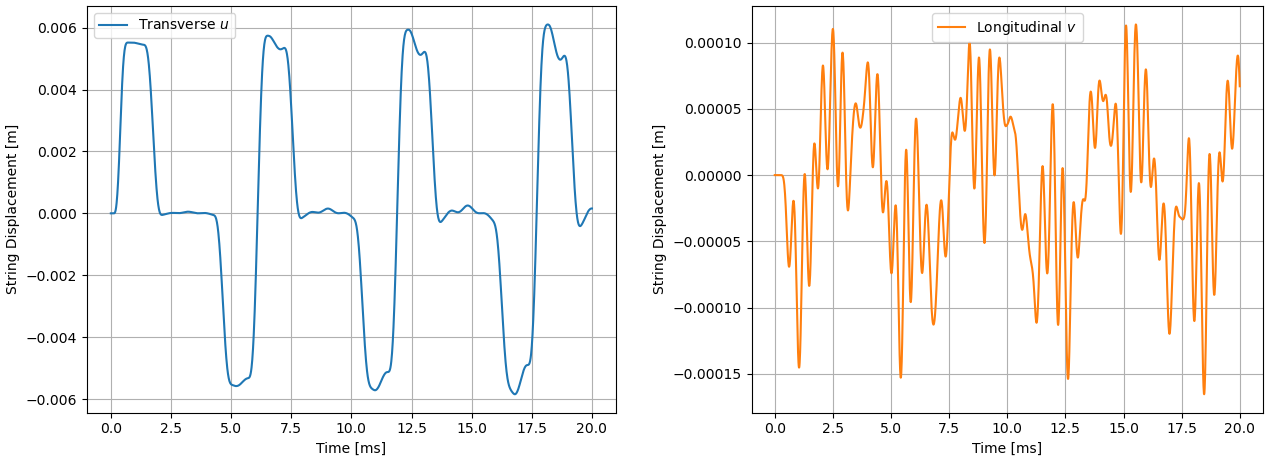

Figure 7: Transverse and Longitudinal displacements of the string at $x = L/4$ computed with SAV.

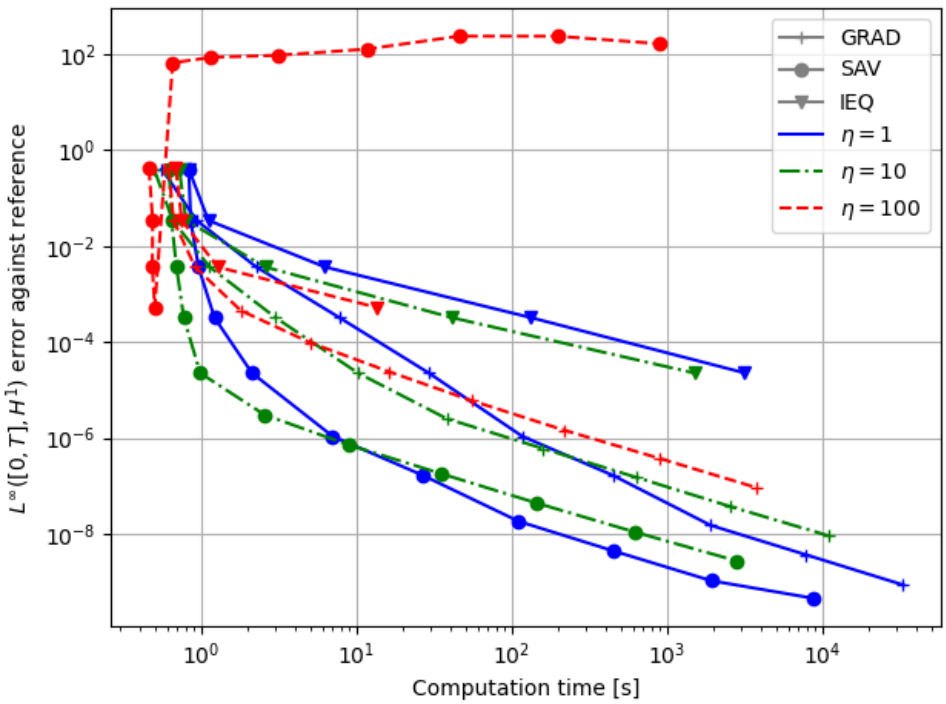

Figure 8: Computation costs of the schemes with respect to accuracy on transverse component u.

Figure 8 shows some computation times and associated errors for 1ms long simulations and for some fixed ratio between the time and space steps Δt and Δx which is

$$\Delta t^2 \rho \left(M_h^{-1} K_h \right) = \eta$$

IEQ is clearly not a good scheme to use in terms of efficiency even with a Woodbury inversion since its cost

increases very rapidly even for large errors.
Compared to the Discrete Gradient with quasi-Newton solver, SAV with Sherman-Morrison is 14 times faster in average to reach a desired precision, but it has some convergence issues that we will investigate in the following sections.

4.2 Energy preservation

The energy balance should be respected with a margin due to machine error on the residuals:

$$\frac{\mathcal{E}_h^{n+1/2} - \mathcal{E}_h^{n-1/2}}{\mathcal{E}_{max}} - F_h^n \cdot \frac{U_h^{n+1} - U_h^{n-1}}{2\mathcal{E}_{max}} = \varepsilon \tag{II.26}$$

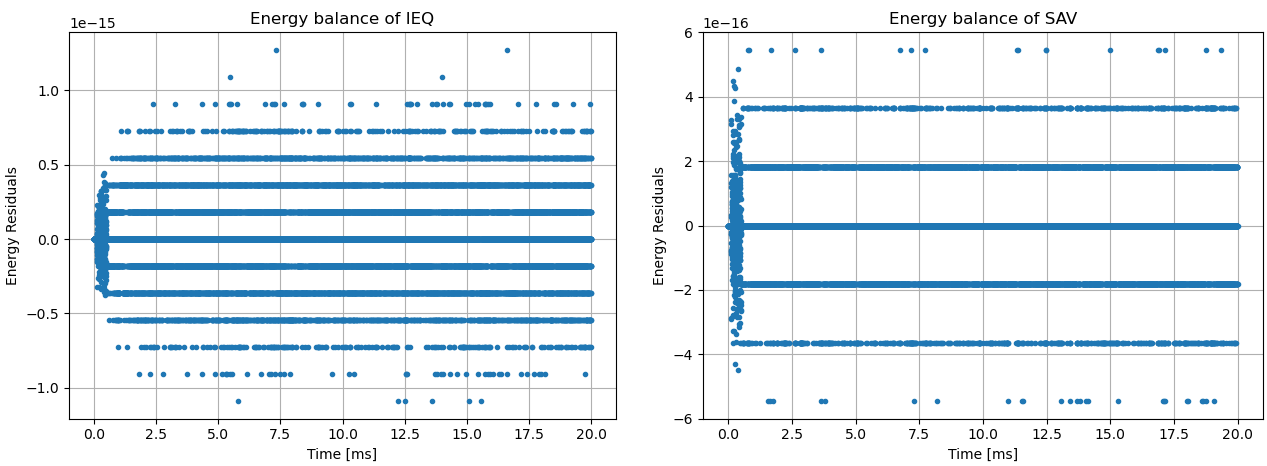

Figure 9: Energy conservation residuals ε of IEQ and SAV schemes.

Both IEQ and SAV schemes show very good energy preservation as illustrated in figure 9.
The noisy part at the beginning of the simulation corresponds to application of the source term. When it stops at $t = 0.5$ms the string oscillates freely and we clearly observe multiples of the machine error on the energy balance's residuals.

4.3 Time convergence

All plots are computed with $T = 1$ ms in the simulations and 10 elements of order 4.
In the following we call "Consecutive $L^\infty([0,T]; H^1)$ error" the error computed between the solution with time step Δt $u_{h,\Delta t}$ and the solution with time step $2\Delta t$ $u_{h,2\Delta t}$:

$$e = \frac{\max\limits_{n \in [\![0,N]\!]} \left\| u_{h,\Delta t}^{2n} - u_{h,2\Delta t}^{n} \right\|_{H^1}}{\max\limits_{n \in [\![0,N]\!]} \left\| u_{h,\Delta t}^{2n} \right\|_{H^1}} \tag{II.27}$$

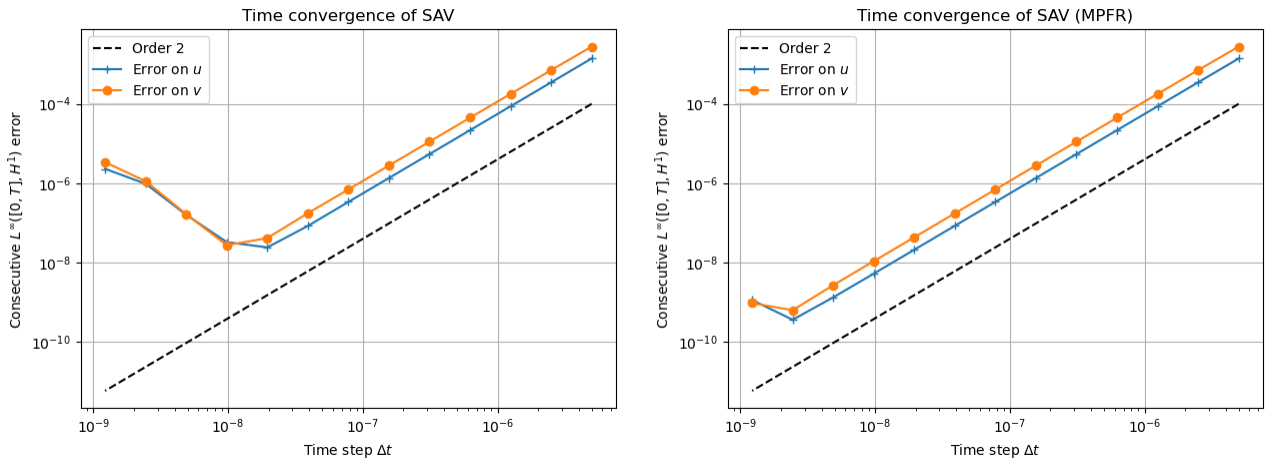

Figure 10: Time convergence of SAV with `'double'` (left) and `'long double'` (right) precision.

The right plot of figure 10 shows convergence curves of SAV scheme with consecutive errors with `'double'` precision. There is an asymptotic behavior for large time steps $\Delta t \in [2 \times 10^{-8}, 5 \times 10^{-6}]$ with a quadratic convergence rate. For small times steps $\Delta t < 2 \times 10^{-8}$ we clearly notice a problem of convergence.
Running the code with a multiple precision library proves that this is an accumulation of numerical errors [Butcher, 2016] and not a problem of convergence, see the right plot in Figure 10.

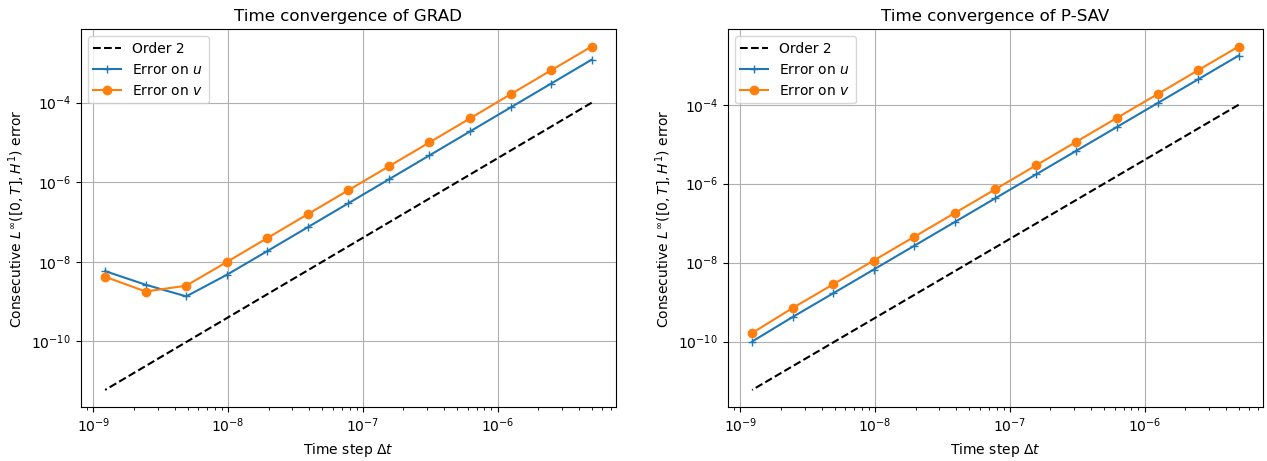

Figure 11: Time convergence of GRAD (left) and P-SAV (right) with `'double'` precision.

Other schemes like Discrete Gradient and P-SAV I.144 show less accumulation of rounding errors for small time steps with `'double'` precision as shown in figure 11.

4.4 Aliasing problems

As mentioned in section 4.1, even if the IEQ and SAV schemes are stable when the CFL is respected (because of theorem 3.1), we observe that the solution can be very polluted and therefore exhibit a very large error with respect to the reference solution.

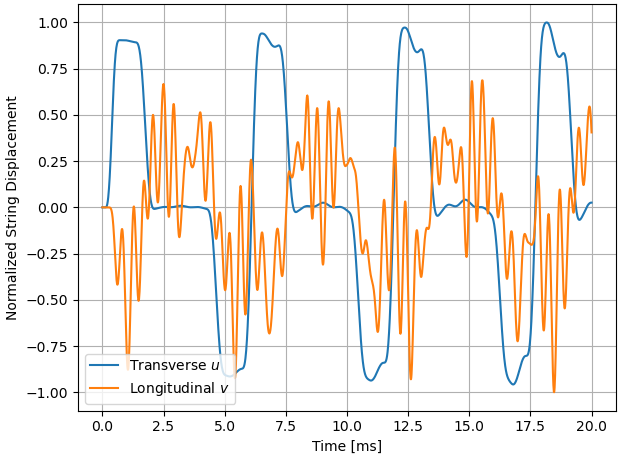

Figure 12: Precise stable simulation.

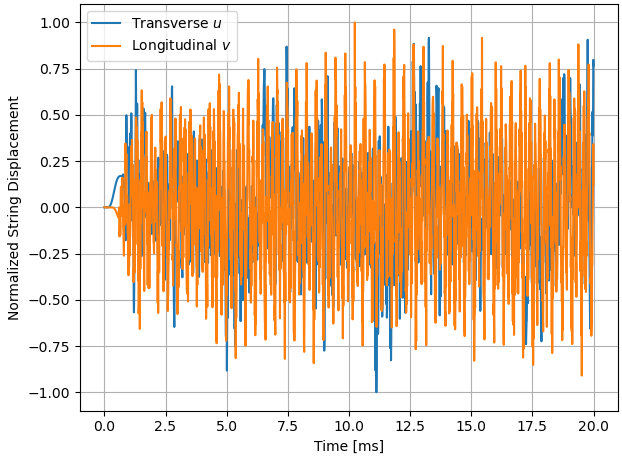

Figure 13: Non-precise stable simulation.

Figures 12 and 13 show two examples of stable simulations, but the one on the left is close to the real solution whereas the one on the right is stable but completely wrong.

We show in the following that these aliasing issues occur when the CFL ratio η is too large, and the threshold can be influenced by the stabilization parameter α.

4.5 Space-Time convergence

Because of the high computational cost of IEQ we present only SAV space-time convergence curves.

In the following we call "$L^\infty([0,T],H^1)$ error against reference" the error computed between a solution u_h and a reference solution $u_{h,\text{ref}}$:

$$e = \frac{\max\limits_{n\in[\![0,N]\!]} \left\| u^n_{h,\Delta t} - u^n_{h,\text{ref}} \right\|_{H^1}}{\max\limits_{n\in[\![0,N]\!]} \left\| u^n_{h,\text{ref}} \right\|_{H^1}} \tag{II.28}$$

4.5.1 Unconditionally stable scheme

The unconditionally stable scheme with $\theta = 1/4$ is used. It has no CFL restriction. The number

$$\eta = \Delta t^2 \rho \left(M_h^{-1} K_h \right) \tag{II.29}$$

is used to choose the time step in relation with a given spatial discretization.
The following figures are computed with `'long double'` precision with $T = 1$ ms simulation duration.

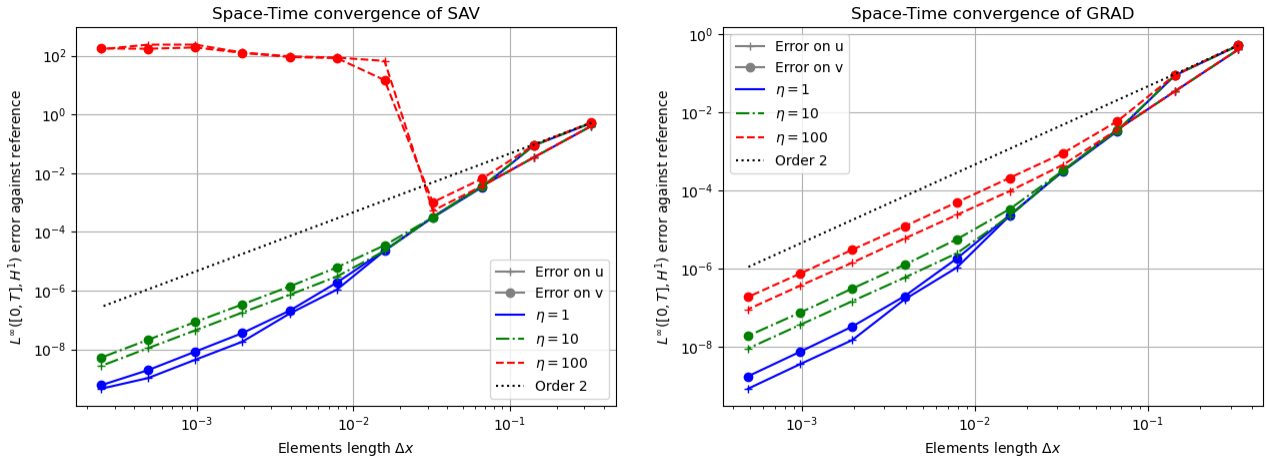

Figure 14: Space-Time convergence of SAV (left) and GRAD (right) schemes.

We observe that unlike the Discrete Gradient (GRAD), the SAV scheme does not always converge with respect to space and time even if it is unconditionally stable. The red curve for $\eta = 100$ on the left of figure 14 is a clear example.
However it seems to converge at order 2 for smaller values of η, at least in this range of spatial discretizations.

We recall that no mathematical result of space-time convergence was obtained in section 3.5 for type 2 nonlinear equations as the piano string. It is still an open question to know whether an extra condition must be fulfilled for space-time convergence or if these schemes simply do not space-time converge at all.

4.5.2 Conditionally stable scheme

To conclude this section we show on figure 15 space-time convergence plots of conditionally stable schemes with $\theta < 1/4$ for which the CFL condition (I.59) must be respected. In this case we denote

$$\eta_\theta = \left(\frac{1}{4} - \theta\right) \Delta t^2 \rho\left(M_h^{-1} K_h\right) = \left(\frac{1}{4} - \theta\right) \eta \tag{II.30}$$

the CFL ratio number that allows to choose the time step from a given spatial discretization.

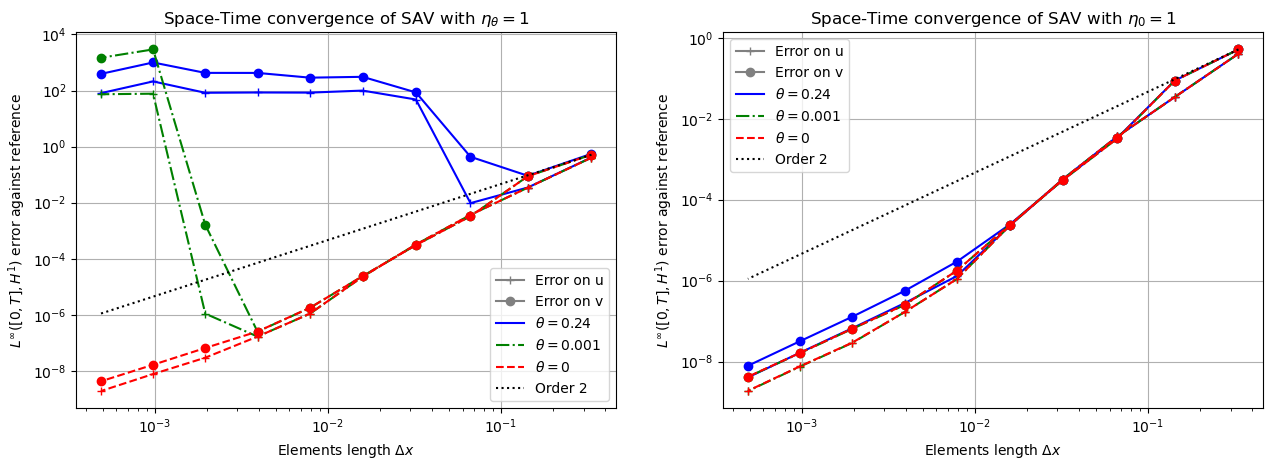

Figure 15: Convergence of CFL-restricted SAV schemes.

First notice that these schemes are stable up to the exact CFL $\left(\frac{1}{4}-\theta\right)\Delta t^2\rho\left(M_h^{-1}K_h\right)=\eta_\theta=1$ and as expected, any attempt to run them with $\eta_\theta>1$ causes the simulations to diverge instantly.
However we clearly see convergence problems occurring on the left plot. On the right plot we ran the same schemes but using the CFL restriction of the explicit scheme $\theta=0$, meaning that the time step used is the one that verifies $\frac{1}{4}\Delta t^2\rho\left(M_h^{-1}K_h\right)=\eta_0=1$. In this case, no more convergence issue is visible.

No clear explanations for these behaviors have been found yet.

4.6 Long time simulations

To verify the long-term precision of the schemes we ran $T=1$s simulations with 10 elements of order 4.

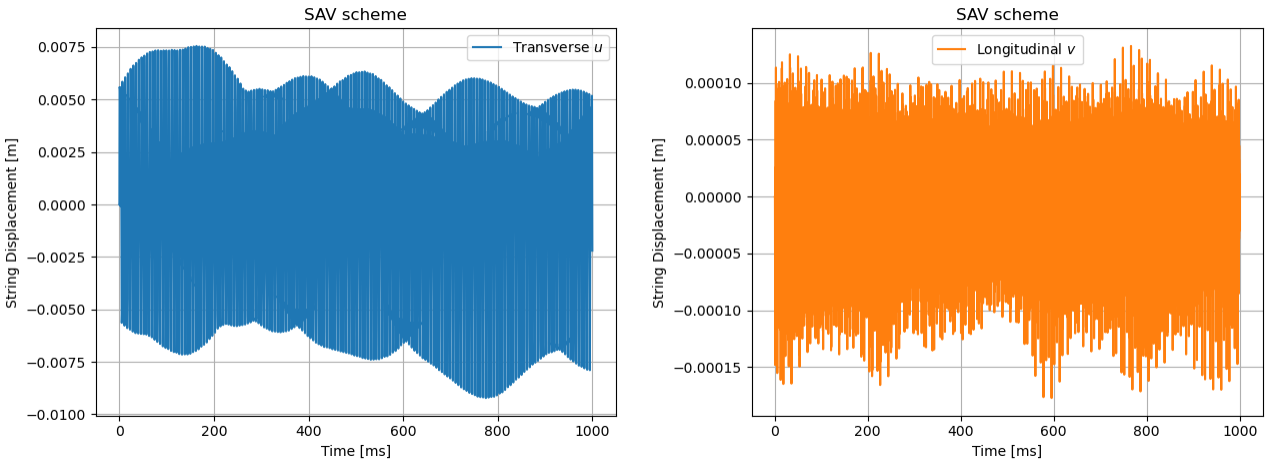

Figure 16: SAV scheme for long time simulation.

SAV remains accurate for long simulations as shown on figure 16. Because of the long simulation time we only distinguish the envelope of the signals which remains in correct range of values.

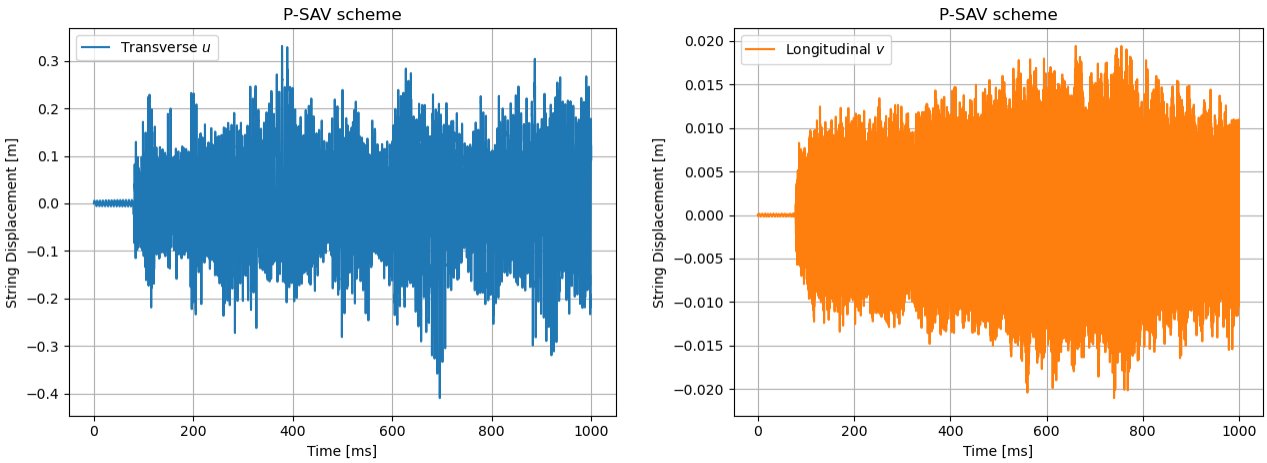

Figure 17: P-SAV scheme for long time simulation.

But P-SAV has a tendency to show aliasing problems after a certain time. Here on figure 17 it is good during the first 100ms of simulation and becomes completely inaccurate after that. The range of values of the solution

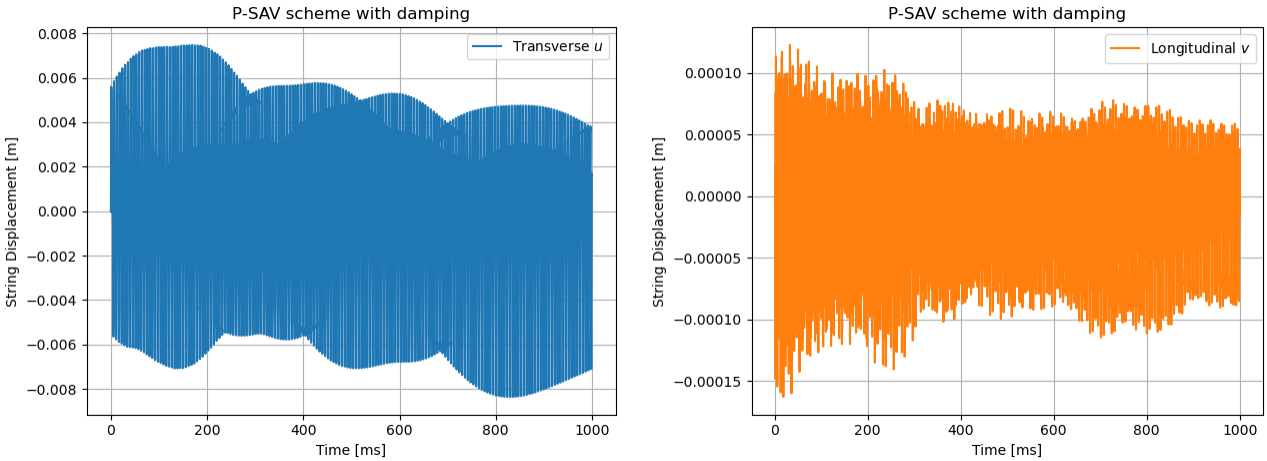

Figure 18: P-SAV scheme for long time simulation with damping.

is a lot too large for this computed case.
However, introducing some physical damping in the string seems to stabilize the scheme, as illustrated in figure 18.
Combined with the fact that P-SAV shows a lot less rounding errors for precise discretizations (see section 4.3) makes it a good candidate for real-case physical simulations.

4.7 Influence of the α-decomposition

As pointed out earlier, the α-decomposition II.8 has an influence on the value of the Lipschitz constant that appears in the convergence estimate of theorem 3.7.

In this section we analyze its influence on the time precision of the SAV scheme, and we remind that this choice has no influence on the complexity and computational cost of the scheme. This is why an appropriate choice is to be done carefully.

4.7.1 Convergence constant estimation

To explore this effect, we computed time convergence consecutive errors and noticed that

$$\begin{aligned} \left\|u_{h,1}^n - u_{h,2}^n\right\|_{L^2} &\leqslant \left\|u_{h,1}^n - u_h(t^n)\right\|_{L^2} + \left\|u_{h,2}^n - u_h(t^n)\right\|_{L^2} && \text{(II.31)} \\ &\leqslant C_\alpha \Delta t_1^2 + C_\alpha \Delta t_2^2 && \text{(II.32)} \\ &\leqslant \frac{5}{4} C_\alpha \Delta t^2 && \text{(II.33)} \end{aligned}$$

if the time steps are divided by 2 between two consecutive simulations.

We use 10 space elements of order 4. We compute 4 different time steps for the convergence curves to determine the convergence constants with an affine regression of $\log(\text{error}) = \log\left(\frac{5}{4}C_\alpha\right) + r\log(\Delta t)$.
We use either large time steps $(10^{-5}, 5\times10^{-6}, 2.5\times10^{-6}, 1.25\times10^{-6})$ or small time steps $(10^{-6}, 5\times10^{-7}, 2.5\times10^{-7}, 1.25\times10^{-7})$.

With this methodology we can deduce the best value of α to ensure the minimal time discretization error.

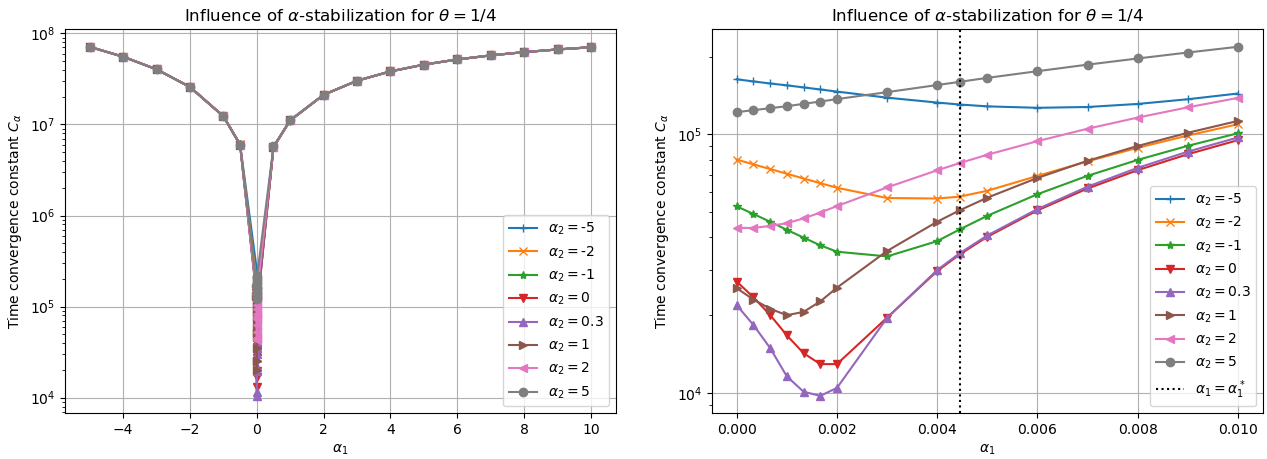

Figure 19: Influence of α for large values of α.

Figure 19 shows the values of the time convergence constants for a wide range of values for α. There is a very clear optimal visible close to $\alpha_1 = 0$ which is presented with a zoom on the right plot. Values of α_2 between 0 and 1 are also the best choices. The following figure 20 focuses on these values.
Also notice that negative values of α actually make the stiffness matrix and the energy negative which is not suitable for a-priori stable simulations.

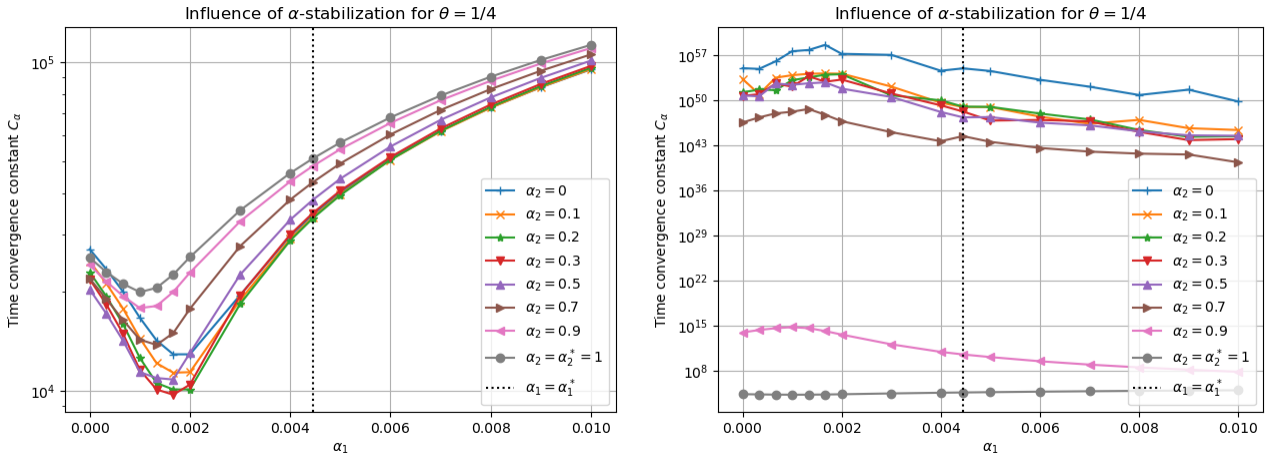

Figure 20: Influence of α for small time steps (left) and large time steps (right).

On the left hand side of figure 20 the times steps are smaller than 10^{-6} and no aliasing is visible. The best value for α_2 seems to be around 0.3 and around 1.5×10^{-3} for α_1.
On the right hand side of figure 20 the time steps are larger that 10^{-6} and aliasing is visible with enormous convergence constants for several values of α. Notice that the more precise the values are on the left, the more aliased they are on the right with large time steps. For example the choice $\alpha_2 = 1$ is the worse choice in terms of precision, but it is the best choice to avoid aliasing.

4.7.2 Lipschitz constant estimation

The Lipschitz constant are evaluated from the computed solutions of the previous section, with the expressions given in proposition 3.4. They are displayed in figure 21 with α^* represented with the dotted vertical line.

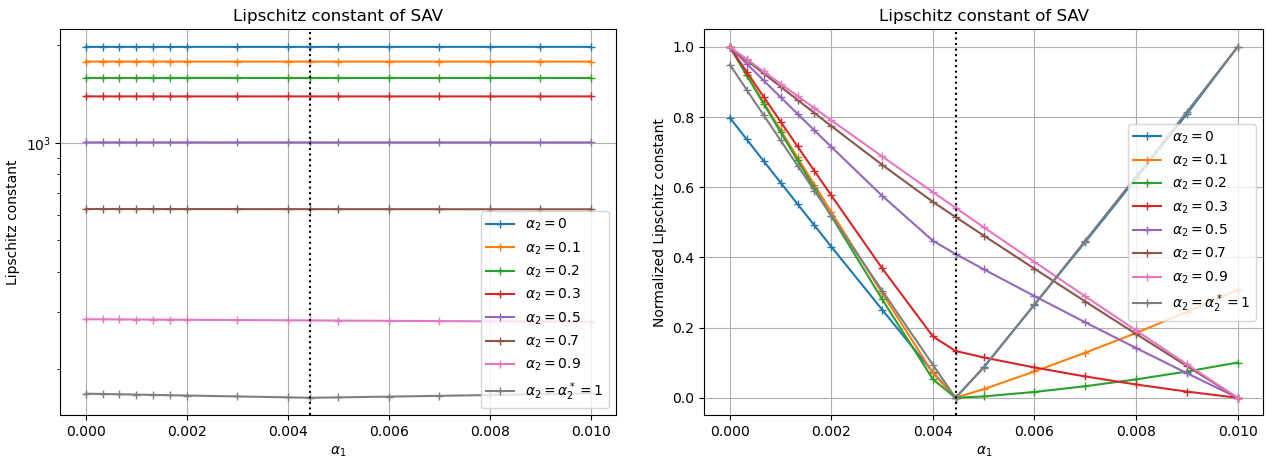

Figure 21: Influence of α on Lipschitz constants.

Figure 21 shows the computed Lipschitz constant corresponding to real simulations. On the left we see that the Lipschitz constant increases when α_2 is close to 0. On the right we normalized these plots the compare the influence of α_1.
The optimal value that makes the Lipschitz constant as small as possible is clearly α^*.

However we saw in the previous paragraph that the optimal value of the time convergence constant is not necessarily α^*.
Moreover, the values of the Lipschitz constant should be exponentially related to the convergence constant of 3.7, which in our present case would give enormous numbers. This may suggest that the estimate of theorem 3.7 is not sharp and that maybe the Lipschitz assumption is not necessary to complete the proof (see [Shen and Xu, 2018] for a proof of space/time convergence without Lipschitz assumption for nonlinear equations of type 1).

4.7.3 Influence of the θ-scheme

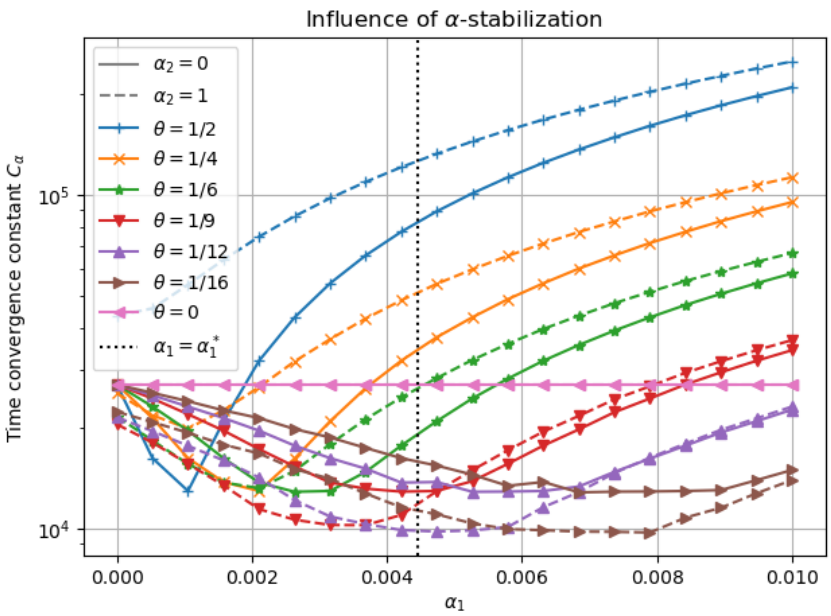

Figure 22: Influence of α for different θ-schemes.

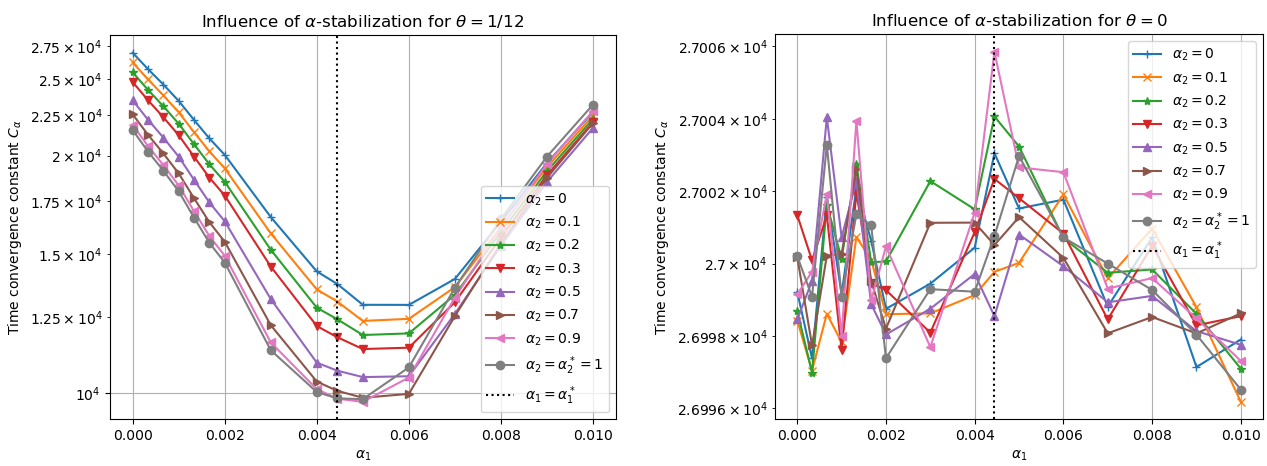

Figure 23: Influence of α for $\theta = 1/12$ (left) and $\theta = 0$ (right).

Figures 22 and 23 shows the time convergence constants for some other values of θ. The optimal value with respect to α is changed. With $\theta = 1/12$ we see that $\alpha_2 = \alpha_2^* = 1$ has become the best choice when it was the

worst one with $\theta = 1/4$.
Figure 22 shows that $\theta = 1/12$ seems to have the best precision associated with $\alpha = \alpha^*$. All the others θ-schemes have their own optimal choices for α.
For $\theta = 0$ (explicit treatment of the linear part) the α-decomposition does no longer have any significant influence on the precision of the scheme.

Notice that the value of α has a very significant influence on the time precision of the schemes for large θ. It can modify the precision by a factor 10. For smaller values of θ is becomes less and less significant.

4.8 Influence of the auxiliary constant c

The choice of a sufficiently large auxiliary constant is mandatory to ensure the good definition of the square root of the auxiliary variable and auxiliary function. A non-suitable choice will cause the scheme to compute NaN very rapidly.
This choice depends on the amplitude of the oscillations and of the source term. Large oscillations require a larger auxiliary constant to offset the larger negative values of the nonlinear function.

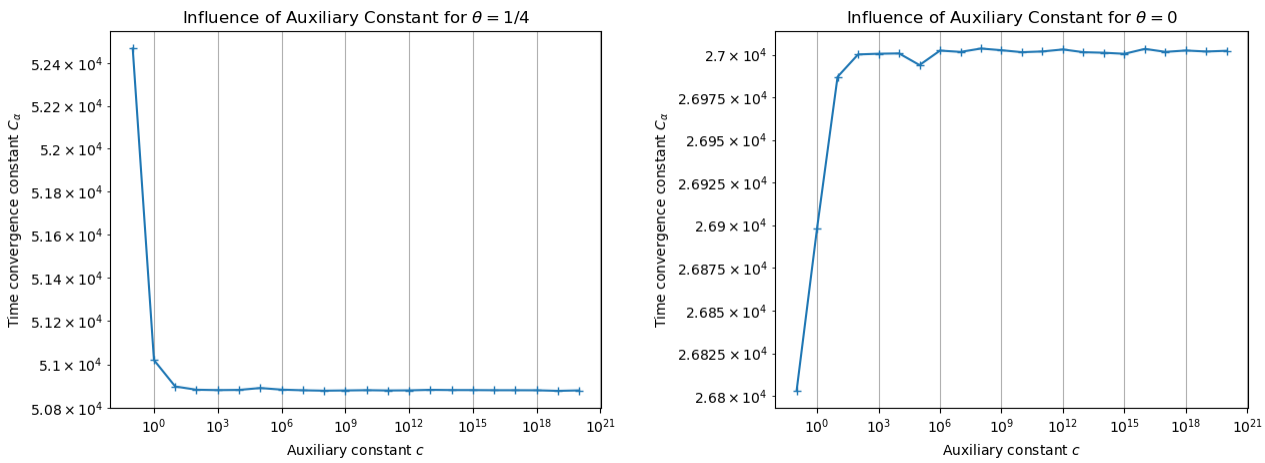

Figure 24: Influence of the auxiliary constant c on the time convergence constant. $\theta = 1/4$ (left), $\theta = 0$ (right).

Following the same process as in section 4.7.1, figure 25 shows the relation between the auxiliary constant c and the time convergence constant for $\theta = 1/4$ on the left and $\theta = 0$ on the right.
We notice that all values give similar results for time precision. Some values appears to be better than others but not very significantly.

Note that unlike the value of α, unfortunate choices of the auxiliary constant c do not cause aliasing.

As before with α, we can compute the Lipschitz constant with respect to the value of c, as displayed in figure 25.

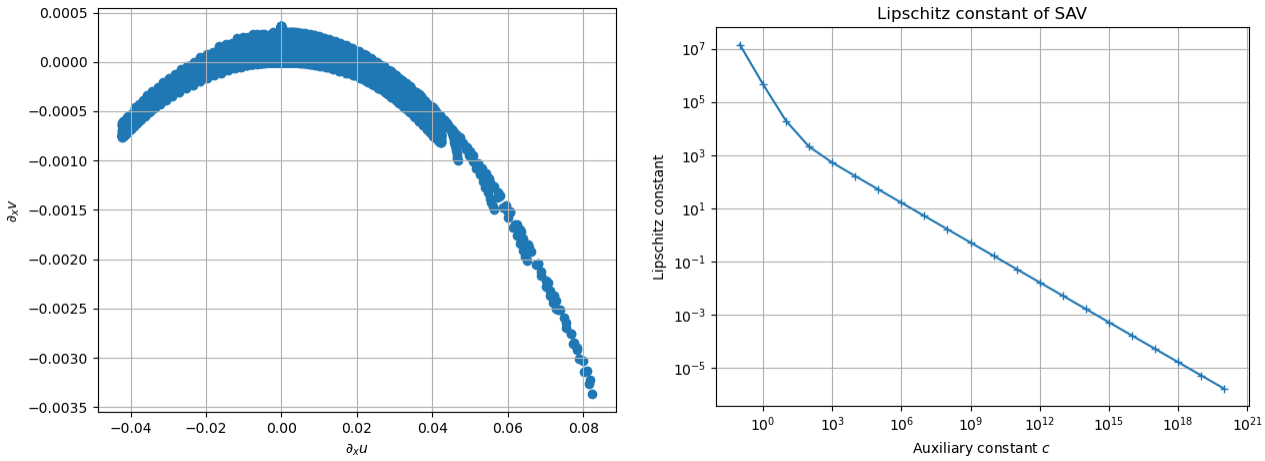

Figure 25: Influence of the auxiliary constant c on the Lipschitz constant.

Once again the effect of the Lipschitz constant on the convergence constant is not visible. The left plot shows an example of set $\mathcal{J}_{\nabla,h}$ of values taken by the gradient of the fully-discrete solution on which the Lipschitz constants are computed.

4.9 Choice of the parameters

In [Ducceschi and Bilbao, 2022, Ducceschi et al., 2022] the choice is made to use $\alpha = \alpha_B = \left(\frac{T_0}{ES}, \frac{T_0}{ES}\right)$ which leads to a particular case of factorization of the nonlinear function:

$$\mathcal{U}_{\alpha_B}(u,v) = \frac{ES - T_0}{2}\left[\sqrt{u^2 + (1+v)^2} - 1\right]^2 \tag{II.34}$$

For the IEQ case [Ducceschi and Bilbao, 2022] it simplifies the auxiliary function and requires no auxiliary constant:

$$g_{\alpha_B}(u,v) = \frac{1}{\sqrt{2\mathcal{U}_{\alpha_B}}}\nabla\mathcal{U}_{\alpha_B} = \sqrt{\frac{ES - T_0}{u^2 + (1+v)^2}}\begin{pmatrix} u \\ 1+v \end{pmatrix} \tag{II.35}$$

However this simplification is not possible for SAV [Ducceschi et al., 2022] because of the integrals in the definition of the auxiliary function.

This value α_B is a correct choice since the $\theta = 0$ scheme is used (see figure 22). But the choice of $\theta = 0$ may not be optimal for precision.

For this specific application case to the piano string we recommend to use the SAV scheme with $(\theta, \alpha, c) = (1/12, \alpha^*, 10^4)$.
If it is used with damping, the P-SAV scheme should be preferred to avoid rounding errors.

For other application cases a similar study should be made, and the benefits of using unconditional schemes or CFL-restricted ones must be evaluated. For example if the problem is very stiff maybe it is better to use $\theta = 1/4$. Or if the use of Sherman-Morrison formula is not possible because of couplings to other systems, the stiffness matrix must be inverted at every time step and it is probably interesting to use $\theta = 0$ instead of $\theta = 1/12$ is terms of computational cost.

5 Conclusions and Prospects

In this report we have presented energy-quadratization techniques that enable us to write linearly implicit and unconditionally stable numerical schemes from Hamiltionian systems of nonlinear wave equations. The Invariant Energy Quadratization (IEQ) technique leads to a scheme of little computational interest compared to discrete gradient methods leading to an iterative solution scheme, due to the large number of degrees of freedom added by the method. The Scalar Auxiliary Variable (SAV) technique, on the other hand, greatly improves performance over an iterative solution technique for a nonlinear problem.

Although very attractive, the mathematical properties of these quadratization methods have been little studied. Proofs of space-time convergence can be found in the literature for type-1 nonlinearities, i.e. nonlinearities that take the solution field as an argument, but no study to our knowledge gives results for type-2 nonlinear terms that take the gradient of the solution field as an argument, as appears, for example, in the geometrically exact piano string.

In the first part of this report, we therefore analyzed the stability and consistency properties of the schemes, then gave an already known proof of convergence for finite differences, which we extended to finite elements and θ-schemas for type 1 nonlinearities. We then highlighted the blocking elements for a proof of space-time convergence in the type 2 case. A proof of convergence in time has also been completed.

The second part is devoted to applications of the quadratized schemes to the piano string model, which exhibits a type 2 nonlinearity.
We find that the quadratic rates of convergence in time agree well with those expected from the theoretical results in Part 1.
We have also noted that the schemes present cases of space-time non-convergence linked to aliasing problems for time steps that are too large, but that, in favorable cases with well-chosen discretization parameters, the schemes still seem to show quadratic space-time convergence. They also show an accumulation of numerical errors, characterized by a rise in errors on the time convergence curves.
The question of the space-time convergence of these schemes for a type-2 nonlinear equation remains open. One approach would be to find an additional convergence condition in addition to the usual CFL stability condition for θ-schemes.

Finally, we studied the influence of discretization parameters on the scheme accuracy, in relation to the convergence bounds obtained in the first part.
We found that the *alpha*-stabilization has a very strong impact on avoiding scheme aliasing, and that it has a significant influence on the discretization error in time. We have put forward stabilization values that give the optimal accuracy of the schemes. These values depend on the θ-schema used. It should also be noted that the stabilization values that give the most accurate results for small time steps are also those most likely to cause aliasing for large time steps.
Although the convergence proofs presented in the first part make use of a Lipschitz-type regularity argument on nonlinear functions, we have noticed that the simulated convergence constants do not show a correlation with the Lipschitz constants as expected, suggesting that the proof could be improved.
Finally, the auxiliary constants used during quadratization appear to have very little influence on the accuracy of the schemes.

Appendices

A Discrete Gronwall Lemma

Lemma A.1 (Modified Discrete Gronwall inequality)
Let $(w_n)_{n\in\mathbb{N}}$ be a non-negative sequence and two non-negative constants $a \in \mathbb{R}_+$ and $c \in \mathbb{R}_+$.
If

$$\forall n \in \mathbb{N}, \quad w_n \leqslant a + c\sum_{i=0}^{n-1}\sum_{j=0}^{i} w_j \tag{II.36}$$

then

$$\forall n \in \mathbb{N}, \quad w_n \leqslant a(1+c)^{\frac{n(n+1)}{2}} \leqslant ae^{\frac{n(n+1)}{2}c} \tag{II.37}$$

Proof *Let's start with the proof of this following result by induction:*

$$1 + \sum_{i=0}^{n-1}\sum_{j=0}^{i} c(1+c)^{\frac{j(j+1)}{2}} \leqslant (1+c)^{\frac{n(n+1)}{2}} \tag{II.38}$$

It holds for $n = 0$ and $n = 1$, so let's suppose that it is true for a specific $n \in \mathbb{N}^*$.

$$\begin{aligned}
1 + \sum_{i=0}^{n}\sum_{j=0}^{i} c(1+c)^{\frac{j(j+1)}{2}} &= 1 + \sum_{i=0}^{n-1}\sum_{j=0}^{i} c(1+c)^{\frac{j(j+1)}{2}} + \sum_{j=0}^{n} c(1+c)^{\frac{j(j+1)}{2}} && \text{(II.39)}\\
&\leqslant (1+c)^{\frac{n(n+1)}{2}} + c\sum_{j=0}^{n}(1+c)^{\frac{j(j+1)}{2}} \quad \textit{because of the induction hypothesis} && \text{(II.40)}\\
&\leqslant (1+c)^{\frac{n(n+1)}{2}} + c(1+c)^{\frac{n(n+1)}{2}}(n+1) \quad \textit{because } c \geqslant 0 && \text{(II.41)}\\
&\leqslant (1+c)^{\frac{n(n+1)}{2}}\,(1+(n+1)c) && \text{(II.42)}\\
&\leqslant (1+c)^{\frac{n(n+1)}{2}}\,(1+c)^{n+1} && \text{(II.43)}\\
&= (1+c)^{\frac{(n+1)(n+2)}{2}} \quad \textit{which concludes the induction.} && \text{(II.44)}
\end{aligned}$$

Now with this result we can prove the lemma by strong induction.
It is easy to check that II.37 is true for $n = 0$ and $n = 1$, so let's suppose that II.37 holds up to rank n and let's prove it at rank $n + 1$:

$$\begin{aligned}
w_{n+1} &\leqslant a + c\sum_{i=0}^{n}\sum_{j=0}^{i} w_j && \text{(II.45)}\\
&\leqslant a + c\sum_{i=0}^{n}\sum_{j=0}^{i} a(1+c)^{\frac{j(j+1)}{2}} \quad \textit{because of the induction hypothesis} && \text{(II.46)}\\
&\leqslant a\left(1 + \sum_{i=0}^{n}\sum_{j=0}^{i} c(1+c)^{\frac{j(j+1)}{2}}\right) && \text{(II.47)}\\
&\leqslant a(1+c)^{\frac{(n+1)(n+2)}{2}} \quad \textit{because of II.38} && \text{(II.48)}
\end{aligned}$$

and the strong induction is completed.
The second inequality in II.37 is true because $1 + c \leqslant e^c$.

References

[Allen and Cahn, 1979] Allen, S. M. and Cahn, J. W. (1979). A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta metallurgica*, 27(6):1085–1095.

[Banks et al., 1995] Banks, H. T., Gaitens, M., Muñoz, B., and Yanyo, L. (1995). Nonlinear elastomers: modeling and estimation. Technical report, North Carolina State University. Center for Research in Scientific Computation.

[Barone et al., 1971] Barone, A., Esposito, F., Magee, C., and Scott, A. (1971). Theory and applications of the sine-gordon equation. *La Rivista del Nuovo Cimento (1971-1977)*, 1(2):227–267.

[Bilbao et al., 2023] Bilbao, S., Ducceschi, M., and Zama, F. (2023). Explicit exactly energy-conserving methods for hamiltonian systems. *Journal of Computational Physics*, 472:111697.

[Bilbao et al., 2015] Bilbao, S., Torin, A., and Chatziioannou, V. (2015). Numerical modeling of collisions in musical instruments. *Acta Acustica united with Acustica*, 101(1):155–173.

[Butcher, 2016] Butcher, J. C. (2016). *Numerical methods for ordinary differential equations*. John Wiley & Sons.

[Cahn and Hilliard, 1958] Cahn, J. W. and Hilliard, J. E. (1958). Free energy of a nonuniform system. i. interfacial free energy. *The Journal of chemical physics*, 28(2):258–267.

[Cai and Shen, 2020] Cai, J. and Shen, J. (2020). Two classes of linearly implicit local energy-preserving approach for general multi-symplectic hamiltonian pdes. *Journal of Computational Physics*, 401:108975.

[Chabassier, 2012] Chabassier, J. (2012). *Modélisation et simulation numérique d'un piano par modèles physiques*. PhD thesis, École polytechnique.

[Chabassier and Imperiale, 2017] Chabassier, J. and Imperiale, S. (2017). Space/time convergence analysis of a class of conservative schemes for linear wave equations. *Comptes Rendus Mathematiques*, 355(3):282–289.

[Chabassier and Imperiale, 2021] Chabassier, J. and Imperiale, S. (2021). Construction and convergence analysis of conservative second order local time discretisation for linear wave equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 55(4):1507–1543.

[Chabassier and Joly, 2010] Chabassier, J. and Joly, P. (2010). Energy preserving schemes for nonlinear hamiltonian systems of wave equations: Application to the vibrating piano string. *Computer Methods in Applied Mechanics and Engineering*, 199(45-48):2779–2795.

[Chatziioannou and Van Walstijn, 2015] Chatziioannou, V. and Van Walstijn, M. (2015). Energy conserving schemes for the simulation of musical instrument contact dynamics. *Journal of Sound and Vibration*, 339:262–279.

[Ducceschi and Bilbao, 2019] Ducceschi, M. and Bilbao, S. (2019). Non-iterative, conservative schemes for geometrically exact nonlinear string vibration.

[Ducceschi and Bilbao, 2022] Ducceschi, M. and Bilbao, S. (2022). Simulation of the geometrically exact nonlinear string via energy quadratisation. *Journal of Sound and Vibration*, 534:117021.

[Ducceschi et al., 2022] Ducceschi, M., Bilbao, S., and Webb, C. J. (2022). Real-time simulation of the struck piano string with geometrically exact nonlinearity via a scalar quadratic energy method. *Mh*, 15:1b.

[Eyre, 1998] Eyre, D. J. (1998). Unconditionally gradient stable time marching the cahn-hilliard equation. *MRS Online Proceedings Library (OPL)*, 529:39.

[Gonzalez, 2000] Gonzalez, O. (2000). Exact energy and momentum conserving algorithms for general models in nonlinear elasticity. *Computer Methods in Applied Mechanics and Engineering*, 190(13-14):1763–1783.

[He and Sun, 2020] He, M. and Sun, P. (2020). Energy-preserving finite element methods for a class of nonlinear wave equations. *Applied Numerical Mathematics*, 157:446–469.

[Jiang et al., 2019] Jiang, C., Cai, W., and Wang, Y. (2019). A linearly implicit and local energy-preserving scheme for the sine-gordon equation based on the invariant energy quadratization approach. *Journal of Scientific Computing*, 80(3):1629–1655.

[Jiang et al., 2021] Jiang, C., Wang, Y., and Gong, Y. (2021). Explicit high-order energy-preserving methods for general hamiltonian partial differential equations. *Journal of Computational and Applied Mathematics*, 388:113298.

[Joly, 2003] Joly, P. (2003). Variational methods for time-dependent wave propagation problems. *Topics in computational wave propagation: direct and inverse problems*, pages 201–264.

[Li and Sun, 2020] Li, D. and Sun, W. (2020). Linearly implicit and high-order energy-conserving schemes for nonlinear wave equations. *Journal of Scientific Computing*, 83:1–17.

[Lin et al., 2019] Lin, L., Yang, Z., and Dong, S. (2019). Numerical approximation of incompressible navier-stokes equations based on an auxiliary energy variable. *Journal of Computational Physics*, 388:1–22.

[Liu and Li, 2020] Liu, Z. and Li, X. (2020). The exponential scalar auxiliary variable (e-sav) approach for phase field models and its explicit computing. *SIAM Journal on Scientific Computing*, 42(3):B630–B655.

[Liu and Li, 2022] Liu, Z. and Li, X. (2022). Step-by-step solving schemes based on scalar auxiliary variable and invariant energy quadratization approaches for gradient flows. *Numerical Algorithms*, pages 1–22.

[Rincon and Quintino, 2016] Rincon, M. A. and Quintino, N. (2016). Numerical analysis and simulation for a nonlinear wave equation. *Journal of Computational and Applied Mathematics*, 296:247–264.

[Rubinstein, 1970] Rubinstein, J. (1970). Sine-gordon equation. *Journal of Mathematical Physics*, 11(1):258–266.

[Shen and Xu, 2018] Shen, J. and Xu, J. (2018). Convergence and error analysis for the scalar auxiliary variable (sav) schemes to gradient flows. *SIAM Journal on Numerical Analysis*, 56(5):2895–2912.

[Shen et al., 2019] Shen, J., Xu, J., and Yang, J. (2019). A new class of efficient and robust energy stable schemes for gradient flows. *SIAM Review*, 61(3):474–506.

[Shen and Yang, 2010] Shen, J. and Yang, X. (2010). Numerical approximations of allen-cahn and cahn-hilliard equations. *Discrete Contin. Dyn. Syst*, 28(4):1669–1691.

[Sherman and Morrison, 1950] Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127.

[Ta-Tsien et al., 1994] Ta-Tsien, L., Yi, Z., and De-Xing, K. (1994). Weak linear degeneracy and global classical solutions for general quasilinear hyperbolic systems. *Communications in partial differential equations*, 19(7-8):1263–1317.

[Woodbury, 1950] Woodbury, M. A. (1950). *Inverting modified matrices*. Statistical Research Group.

[Yang, 2016] Yang, X. (2016). Linear, first and second-order, unconditionally energy stable numerical schemes for the phase field model of homopolymer blends. *Journal of Computational Physics*, 327:294–316.

[Yang and Zhang, 2020] Yang, X. and Zhang, G.-D. (2020). Convergence analysis for the invariant energy quadratization (ieq) schemes for solving the cahn–hilliard and allen–cahn equations with general nonlinear potential. *Journal of scientific computing*, 82:1–28.

[Zhao et al., 2017] Zhao, J., Wang, Q., and Yang, X. (2017). Numerical approximations for a phase field dendritic crystal growth model based on the invariant energy quadratization approach. *International Journal for Numerical Methods in Engineering*, 110(3):279–300.

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 avenue de la Vieille Tour
33405 Talence Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399