



**HAL**  
open science

# High order schemes for hyperbolic problems using continuous finite elements and avoiding mass matrices.

Remi Abgrall

► **To cite this version:**

Remi Abgrall. High order schemes for hyperbolic problems using continuous finite elements and avoiding mass matrices.. 2017. hal-01445543v1

**HAL Id: hal-01445543**

**<https://hal.science/hal-01445543v1>**

Preprint submitted on 25 Jan 2017 (v1), last revised 8 Jul 2017 (v2)

**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.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

# High order schemes for hyperbolic problems using continuous finite elements and avoiding mass matrices.

R. Abgrall  
Institute of Mathematics, University of Zurich  
CH 8057 Zurich, Switzerland

January 25, 2017

## Abstract

When integrating unsteady problems with continuous finite element methods, one faces the problem of inverting a mass matrix. In some cases, one has to recompute this mass matrix at each time steps, in some methods that are not directly formulated by standard variational principles, it is not clear how to write an invertible mass matrix. Hence, in this paper, we show how to avoid this problem for hyperbolic systems, and we also detail the conditions under which this is possible. Analysis and simulation support our conclusions, namely that it is possible to avoid inverting mass matrices without sacrificing the accuracy of the scheme. This paper is an extension of [3] and [19].

## 1 Introduction

We are interested in the numerical approximation of the hyperbolic problem

$$\frac{\partial u}{\partial t} + \operatorname{div} \mathbf{f}(\mathbf{u}) = 0 \quad \mathbf{x} \in \Omega \subset \mathbb{R}^d \quad (1)$$

with the initial condition and boundary conditions, by mean of a finite element like technique. More precisely, the physical space is covered by a conformal tessellation  $\mathcal{T}$ . For ease of exposure, we assume that

$$\Omega = \cup_{K \in \mathcal{T}} K.$$

The solution of the problem is approximated in the space  $V^h$  defined by:

$$V^h = \{\mathbf{u}^h \in C^0(\Omega) \text{ such that for any } K, \mathbf{u}|_K \text{ is a polynomial of degree } r\}.$$

We denote by  $\mathbb{P}_r$  the set of polynomials of degree  $r$ .

It is well known that any finite element technique applied to (1) will lead to a formulation of the type

$$M \frac{dU}{dt} + F = 0$$

where  $U$  denotes the vector of degrees of freedom,  $F$  is an approximation of  $\operatorname{div} \mathbf{f}$  and  $M$  is a mass matrix. In the case of continuous elements, this matrix is sparse but is not block diagonal, contrarily to what happens for the Discontinuous Galerkin methods where the global continuity

requirement is not made. Hence, in order to use any standard ODE solver, we need to invert  $M$ . This is considered by cumbersome by many practitioners and this has been, in our opinion, one of the factors that has led to supremacy of DG methods in the current development of high order schemes. Another drawback appears when we need to reconstruct frequently the mass matrix, such as in ALE algorithms, or by using the SUPG method. In some cases, we do not even know if the mass matrix is invertible, nor even what is the clear variational formulation of the discrete problem. This seemingly strange behavior comes from the fact that the scheme may only have a purely discrete formulation, and then one can come up to a variational formulation that may not be unique, see [1], for one example.

Hence there is a need to construct time integrators for problems of the type (1) that does not require the inversion of a mass matrix while the spatial approximation is done via a continuous element. However, practitioners have often spent years in designing their spatial approximation, and it is out of question to modify this. This paper is precisely trying to give an answer to this apparent contradiction, and to do so we will provide several examples.

The rest of the paper is organized as follows. First, we detail our version of problem (1), and recall the notion of weak solutions. In a second section, we make a short presentation of Defect-Correction (DeC) time integrators that will be at the core of our method. The fourth section presents a detailed description of our method, as well as its analysis. The fifth section presents numerical tests, both for linear and non linear problems, with several space integrators. Accuracy and stability are tested extensively. A conclusion follows. In the appendix, we present several variants of the method that can be of interest, for example when wishes other time integrators.

In the text,  $C$  is any constant, and we apply the usual rule stating that  $C \times \alpha \rightarrow C$  for any positive  $\alpha \in \mathbb{R}^+$ .

## 2 Continuous problem setting

We consider

$$\frac{\partial u}{\partial t} + \operatorname{div} \mathbf{f}(\mathbf{u}) = 0 \quad \mathbf{x} \in \Omega \subset \mathbb{R}^d \quad (2a)$$

with the initial condition and boundary conditions. The initial condition is

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}) \quad \mathbf{x} \in \Omega, \quad (2b)$$

and we also consider boundary conditions on the inflow boundary. Let us give a precise meaning to this. In order to impose a condition  $u = g$  on the inflow boundary, we assume, for any real  $a$  and  $b$ , and any vector  $\mathbf{n}$ , the existence of  $\overline{\nabla_u \mathbf{f}}(a, b)$  such that

- $\overline{\nabla_u \mathbf{f}}(a, a) = \nabla_u \mathbf{f}(a) \cdot \mathbf{n}$  and,
- $\mathbf{f}(b) \cdot \mathbf{n} - \mathbf{f}(a) \cdot \mathbf{n} = \overline{\nabla_u \mathbf{f}}(a, b) (b - a)$ .

This is a reminiscence of the Roe average. As soon as  $\mathbf{f}$  is  $C^1$ , such average exists and is unique. Then the boundary conditions are set such that

$$\max(0, \overline{\nabla_u \mathbf{f}}(\mathbf{u}, g) \cdot \mathbf{n})(\mathbf{u} - g) = 0 \text{ on } \Gamma. \quad (2c)$$

We introduce the flux  $\mathcal{F}$  defined by:

$$\mathcal{F}(a, b) = \frac{1}{2} \left( \mathbf{f}(a) \cdot \mathbf{n} + \mathbf{f}(b) \cdot \mathbf{n} + |\overline{\nabla_u \mathbf{f}}|(a, b) \right).$$

We introduce the space

$$C_{0,t}^1(\Omega \times \mathbb{R}^+) = \{\varphi \in C^1(\Omega \times \mathbb{R}^+) \text{ such that there exists } T_u \text{ for which for any } \mathbf{x} \in \Omega, \\ u(\mathbf{x}, t) = 0 \text{ when } t > T_u\}.$$

The weak form of (2) is: find  $\mathbf{u} \in \mathcal{L}^1(\Omega \times \mathbb{R}^+) \cap L_{loc}^\infty(\Omega \times \mathbb{R}^+)$  such that for any  $\varphi \in C_{0,t}^1(\Omega \times \mathbb{R}^+)$

$$\int_{\Omega \times \mathbb{R}^+} \nabla \varphi \cdot \mathbf{f}(u) d\mathbf{x} dt + \int_{\Omega} \varphi(\mathbf{x}, 0) u_0(\mathbf{x}) d\mathbf{x} - \int_{\mathbb{R}^+} \int_{\Gamma} \varphi(\mathbf{x}, t) (\mathbf{f}(\mathbf{u}) \cdot \mathbf{n} - \mathcal{F}(\mathbf{u}, g, \mathbf{n})) d\omega = 0 \quad (3)$$

This is a direct generalization of what is done for  $\mathbf{f}(\mathbf{u}) = \beta \mathbf{u}$  when  $\beta$  is constant. See [7] for more. For existence and uniqueness results, see for example [12].

### 3 Short review of DeC schemes for ODEs

We consider the problem:

$$\begin{aligned} \frac{dy}{dt} &= f(y, t) \\ y(0) &= y_0 \end{aligned} \quad (4)$$

with suitable conditions on  $f$  to guaranty existence and uniqueness of the solution. There exists a considerable number of methods to solve numerically (4): Runge-Kutta methods, multistep methods, etc. They can be explicit, implicit or semi-implicit. In the case of implicit methods, there are several options depending on where the user wants to put the complexity. One of the strategies is to use a Differed Correction scheme. Let us write formally the (discrete) problem to solve as  $\mathcal{L}^2 = 0$ ; it is a complex one because it is implicit and in general  $f$  in (4) is non linear. Assume we have a cheap way to solve it by a method written formally as  $\mathcal{L}^1 = 0$ . Since it is cheap, the results can be of poor quality. The idea is to use the solution problems  $\mathcal{L}^1 = b$ , for  $b$  well chosen, to construct a sequence of approximations of  $y$ . In DeC, the choice of  $b$  enable to make a controlled and small number of iterations, so that the final term of the sequence is an accurate approximation of  $\mathcal{L}^2 = 0$ . The aim of this section is to make this series of statements more precise. These schemes have recently been revisited by [11, 8, 18], see also the references therein.

We know that the solution of (4) satisfies:

$$y(t) = y(0) + \int_0^t f(y(s), s) ds.$$

The idea is to mimic as much as possible the Picard iteration:

$$y^{n+1}(t) = y(0) + \int_0^t f(y^n(s), s) ds \quad (5)$$

which is known to converge if  $t$  is small enough (and related to the maximum (in  $t$ ) Lipschitz constant of  $f(\cdot, t)$ ).

Let us be given a time step  $\Delta t$  and we define as usual  $t_k = k\Delta t$ ,  $k \in \mathbb{N}$ . It is possible to mimic the Picard iterations (5) on  $[t_n, t_{n+1}]$  by using the following procedure:

- We choose interpolation points in  $[t_n, t_{n+1}]$ , namely:  $t_{n,\ell} = t_n + \xi_\ell \Delta t$ ,  $\ell = 0, \dots, r$  and  $0 = \xi_0 < \dots < \xi_\ell < \xi_{\ell+1} < \dots < \xi_r = 1$ . In particular,  $t_{n,0} = t_n$  and  $t_{n,r} = t_{n+1}$ . An approximation of  $y(t_{n,\ell})$  is denoted by  $y_\ell$ ,  $\ell = 0, \dots, r$

- We define the forward method as : for any  $0 \leq \ell \leq r - 1$ ,

$$y_{\ell+1} = y_{\ell} + \xi_{\ell} \Delta t f(y_{\ell}, t_{n,\ell}). \quad (6)$$

This is the forward Euler method.

We introduce the vector  $(y_1, \dots, y_{r+1})^T$  solution of

$$\mathcal{L}^1(y_1, \dots, y_r) = 0 \quad (7)$$

where  $\mathcal{L}^1$  is defined by

$$\mathcal{L}^1(y_1, \dots, y_r) = \begin{pmatrix} y_r - y_{r-1} - \xi_{r-1} \Delta t f(y_{r-1}, t_{n,r-1}) \\ \vdots \\ y_1 - y_0 - \xi_0 \Delta t f(y_0, t_{n,0}) \end{pmatrix}. \quad (8)$$

Here,  $y_0 \approx y(t_n)$ . Introducing  $\mathcal{I}_0$  the piecewise constant interpolant of  $(f(y_0, t_{n,0}), \dots, f(y_{r+1}, t_{n,r+1}))$ : for any  $s \in [t_{n,0}, t_{n,r}] = [t_n, t_{n+1}]$ , define

$$\text{if } s \in [t_{n,l}, t_{n,l+1}[ \text{ for } 0 \leq l \leq r - 1, \mathcal{I}_0(f(y_0, t_{n,0}), \dots, f(y_{r+1}, t_{n,r+1}); s) = f(y_l, t_{n,l}),$$

we can rewrite (8) as

$$\mathcal{L}^1(y_1, \dots, y_{r+1}) = \begin{pmatrix} y_r - y_0 - \Delta t \int_{t_{n,0}}^{t_{n,r}} \mathcal{I}_0(f(y_0, t_{n,0}), \dots, f(y_{r+1}, t_{n,r+1}); s) ds \\ \vdots \\ y_1 - y_0 - \Delta t \int_{t_{n,0}}^{t_{n,1}} \mathcal{I}_0(f(y_0, t_{n,0}), \dots, f(y_{r+1}, t_{n,r+1}); s) ds \end{pmatrix} \quad (9)$$

- Define  $\mathcal{L}^2$  by:

$$\mathcal{L}^2(y_1, \dots, y_r) = \begin{pmatrix} y_r - y_0 - \Delta t \int_{t_{n,0}}^{t_{n,r}} \mathcal{I}_r(f(y_0, t_{n,0}), \dots, f(y_r, t_{n,r}); s) ds \\ \vdots \\ y_1 - y_0 - \Delta t \int_{t_{n,0}}^{t_{n,1}} \mathcal{I}_r(f(y_0, t_{n,0}), \dots, f(y_r, t_{n,r}); s) ds \end{pmatrix} \quad (10)$$

where  $\mathcal{I}_r$  is the interpolant of degree  $r$  at the  $t_{n,l}$  of the  $f(y_l, t_{n,l})$ . Clearly there exists real  $\theta_{l,m}$  such that

$$\int_{t_{n,0}}^{t_{n,m}} \mathcal{I}_r(f(y_0, t_{n,0}), \dots, f(y_r, t_{n,r}); s) ds = \sum_{l=0}^r \theta_{l,m} f(y_l, t_{n,l})$$

so that we can rewrite (10) as:

$$\mathcal{L}^2(y) = \begin{pmatrix} y_r - y_0 - \Delta t \sum_{l=1}^{r+1} \theta_{l,r} f(y_l, t_{n,l}) \\ \vdots \\ y_1 - y_0 - \Delta t \sum_{l=1}^{r+1} \theta_{l,1} f(y_l, t_{n,l}) \end{pmatrix} \quad (11)$$

The method is defined as follows: We are given  $y^0 = y(0)$ , then we construct  $\{y^k\}_{k \geq 0}$  by induction:

1. Start from  $y^{(0)} = (y^n, \dots, y^n)^T$  where  $y^n \approx y(t_n)$ . This means that  $y_0 = y^n$ .
2. Define  $y^{(1)}$  as the solution of  $\mathcal{L}^1(y^{(1)}) = 0$
3. for  $m = 1, \dots, M$ , define  $y^{(m+1)}$  as the solution of:

$$\mathcal{L}^1(y^{(m+1)}) = \mathcal{L}^1(y^{(m)}) - \mathcal{L}^2(y^{(m)}).$$

4. Set  $y^{n+1} = y_{r+1}^{(M)}$ .

In practice, provided conditions recalled in section 4.2.1 are met, at most  $M$  iterations are needed to have a  $\min(M + 1, r + 1)$ th order accurate approximation of the solution.

We have a variant of this method by replacing the operator  $\mathcal{L}^1$  by:

$$\mathcal{L}^1(y) = \begin{pmatrix} y_r - y_0 - \alpha_r \Delta t f(y_0, t_{n,0}) \\ \vdots \\ y_1 - y_0 - \alpha_1 \Delta t f(y_0, t_{n,0}) \end{pmatrix} \quad (12)$$

with  $\alpha_l = \sum_{j=1}^l \xi_j$ .

**Remark 3.1** (Notations.). *In order to simplify the notations, if  $f$  is a function  $g$  defined on  $\mathbb{R}$ , and  $I_p$  its lagrange interpolant at some stencil of  $p + 1$  distinct points  $t_0 < t_1 < \dots < t_{p+1}$ . Instead of writing  $I_p(g(t_0), \dots, g(t_{p+1}))$ , we write in the sequel  $I_p(g)$  since there will be no ambiguity on the stenci. For example (10) writes:*

$$\mathcal{L}^2(y_1, \dots, y_r) = \begin{pmatrix} y_r - y_0 - \Delta t \int_{t_{n,0}}^{t_{n,r}} \mathcal{I}_r(f; s) ds \\ \vdots \\ y_1 - y_0 - \Delta t \int_{t_{n,0}}^{t_{n,1}} \mathcal{I}_r(f; s) ds \end{pmatrix}$$

## 4 Application to the convection problem

We are interested in solving

$$\frac{\partial u}{\partial t} + \operatorname{div} \mathbf{f}(u) = 0 \quad (13)$$

subjected to  $u(x, 0) = u_0(x)$ . We assume to work in  $\mathbb{R}^2$  to simplify the text. The case with boundary conditions is similar. We are given a triangulation of  $\mathbb{R}^d$ . Here  $d = 2$ , but the discussion is general. The elements are denoted by  $K$  and are (assumed to be) convex. In each element, we assume that the solution is approximated by a polynomial of degree  $r$  and that the approximation is globally continuous. Let us denote this by  $u^h$ . The function  $u^h$  is fully defined by its control parameter  $u_\sigma$  at all the degrees of freedom  $\sigma$ . We define by  $\mathcal{S}$  the set of degrees of freedom, so that

$$u^h = \sum_{\sigma \in \mathcal{S}} u_\sigma \varphi_\sigma.$$

We denote by  $V_h = \text{span} \{\varphi_\sigma, \sigma \in \mathcal{S}\}$ . For now on, we can think of  $u_\sigma$  as the value of  $u^h$  at  $\sigma$  and thus as the  $\varphi_\sigma$  as the Lagrange basis, but we will need slightly less conventional approximations later. Note that DeC time stepping methods have already been used for convection dominated problems, see for example [17].

We assume that we have a good integrator of the steady version of (13), and that this scheme writes: for any degree of freedom  $\sigma$ ,  $u^h$  satisfies:

$$\sum_{K \ni \sigma} \Phi_\sigma^{K,\mathbf{x}}(u^h) = 0. \quad (14)$$

Examples are given by:

- The SUPG residual [13]:

$$\Phi_\sigma^{\mathbf{x}}(u^h) = \int_{\partial K} \varphi_\sigma \mathbf{f}(u^h) \cdot \mathbf{n} - \int_K \nabla \varphi_\sigma \cdot \mathbf{f}(u^h) + h_K \int_K \left( \nabla_u \mathbf{f}(u^h) \cdot \nabla \varphi_\sigma \right) \tau \left( \nabla_u \mathbf{f}(u^h) \cdot \nabla u^h \right) \quad (15)$$

with  $\tau > 0$ .

- The Galerkin scheme with jump stabilization, see [9] for details:

$$\Phi_\sigma^{\mathbf{x}}(u^h) = \int_{\partial K} \varphi_\sigma \mathbf{f}(u^h) \cdot \mathbf{n} - \int_K \nabla \varphi_\sigma \cdot \mathbf{f}(u^h) + \sum_{\text{edges of } K} \Gamma h_e^2 \int_e [\nabla u] \cdot [\nabla \varphi_\sigma] \quad (16)$$

with  $\Gamma > 0$ . Here, since the mesh is conformal, any edge  $e$  (or face in 3D) is the intersection of the element  $K$  and an other element denoted by  $K^+$ . For any function  $\psi$ , we define  $[\nabla \psi] = \nabla \psi|_K - \nabla \psi|_{K^+}$ .

- The limited Residual distributive scheme (RDS), see [5, 4, 2], namely

$$\Phi_\sigma^{\mathbf{x}} = \beta_\sigma \int_{\partial K} \mathbf{f}(u^h) \cdot \mathbf{n} + h_K \int_K \left( \nabla_u \mathbf{f}(u^h) \cdot \nabla \varphi_\sigma \right) \tau \left( \nabla_u \mathbf{f}(u^h) \cdot \nabla u^h \right) \quad (17)$$

or

$$\Phi_\sigma^{\mathbf{x}}(u^h) = \beta_\sigma \int_{\partial K} \mathbf{f}(u^h) \cdot \mathbf{n} + \sum_{\text{edges of } K} \Gamma h_e^2 \int_e [\nabla u] \cdot [\nabla \varphi_\sigma]. \quad (18)$$

We notice that the SUPG and the limited RDS residuals write as

$$\Phi_\sigma^K = \int_K \psi_\sigma \text{div } \mathbf{f}(u) dx$$

where:

- $\Psi_\sigma = \varphi_\sigma + h_K \left( \nabla_u \mathbf{f}(u^h) \cdot \nabla \varphi_\sigma \right) \tau$  for the SUPG scheme,
- for the limited RDS (15), we take

$$\psi_\sigma = \beta_\sigma + h_K \left( \nabla_u \mathbf{f}(u^h) \cdot \nabla \varphi_\sigma \right) \tau.$$

This implies that formally at least, the exact solution cancels the SUPG and RDS residuals. In the case of the stabilisation by jumps (18), we can only write that

$$\Phi_\sigma^K = \int_K \psi_\sigma \operatorname{div} \mathbf{f}(u) dx + R_\sigma(u^h)$$

with

$$R_\sigma = \sum_{\text{edges of } K} \Gamma h_e^2 \int_e [\nabla u] \cdot [\nabla \varphi_\sigma].$$

We note that

- for the RDS scheme (18),  $(\psi_\sigma)|_K = \beta_\sigma^K$
- and for the Galerkin scheme stabilized by jumps,
- $\sum_{\sigma \in K} R_\sigma = 0$ .

The additional term  $R_\sigma$  is not zero, except for the exact solution unless this solution has continuous normal gradients, see [9] for more details. In any case, we note that

$$\sum_{\sigma \in K} \psi_\sigma = 1. \tag{19}$$

**Remark 4.1** (Variational formulation). *It is well known that the SUPG and Galerkin method with jump stabilisation emanate from a variational formulation, namely (and forgetting the boundary conditions to avoid introduce additional notations at this point)*

- For SUPG

$$a(u, v) = - \int_\Omega \nabla v \cdot f(u) dx + \sum_K h_K \int_K \nabla f(u^h) \cdot \nabla v^h \tau \nabla f(u^h) \cdot \nabla x^h dx + \text{Boundary conditions.}$$

- For Galerkin with jump stabilisation

$$a(u, v) = - \int_\Omega \nabla v \cdot f(u) dx + \sum_{\text{internal faces}} h_e^2 \Gamma_e \int_e [\nabla u^h] [\nabla v^h] + \text{Boundary conditions.}$$

*this is not the case of the RDS formulations in general, see [1].*

#### 4.1 Formulation for unsteady problems

Similar to the ODE problem, we could integrate (13) in time and get:

$$u(\mathbf{x}, t) = u(\mathbf{x}, 0) + \int_0^t \operatorname{div} \mathbf{f}(u(x, s)) ds,$$

and the approximate by

$$u(\mathbf{x}, t) \approx u(\mathbf{x}, 0) + t \sum_{l=0}^r \omega_l \operatorname{div} \mathbf{f}(u(\mathbf{x}, s_l)) ds,$$

this with the same conventions as in the ODE case. This suggests the algorithm we describe now where  $V_0$  plays the role of  $u^n$  and is *fixed*. For any  $V \in V_h$ ,  $V^\sigma$  is the control parameter at the degree of freedom  $\sigma$ :  $V = \sum_{\sigma \in \mathcal{S}} V^\sigma \varphi_\sigma$ . Then,



- For any  $\sigma \in \mathcal{S}$ , define  $\mathcal{L}_\sigma^1$  as:

$$\mathcal{L}_\sigma^1(V_1, \dots, V_r) = \begin{pmatrix} |C_\sigma|(V_r^\sigma - V_{r-1}^\sigma) + \sum_{K \ni \sigma} \int_{t_{n,0}}^{t_{n,r}} \mathcal{I}_0(\Phi_\sigma^x(V), s) ds \\ \vdots \\ |C_\sigma|(V_1^\sigma - V_0^\sigma) + \sum_{K \ni \sigma} \int_{t_{n,0}}^{t_{n,1}} \mathcal{I}_0(\Phi_\sigma^x(V), s) ds \end{pmatrix} \quad (20a)$$

The quantity  $|C_\sigma|$ , which plays the same role as the measure of a dual cell, will be defined later in the text.

- and define  $\mathcal{L}_\sigma^2$  by

$$\mathcal{L}_\sigma^2(V_1, \dots, V_r) = \begin{pmatrix} \sum_{K \ni \sigma} \left( \int_K \Psi_\sigma(V_r - V_0) dx + \int_{t_{n,0}}^{t_{n,r}} \mathcal{I}_{r+1}(\Phi_\sigma^x(V), s) ds \right) \\ \vdots \\ \sum_{K \ni \sigma} \left( \int_K \Psi_\sigma(V_1 - V_0) dx + \int_{t_{n,0}}^{t_{n,1}} \mathcal{I}_{r+1}(\Phi_\sigma^x(V), s) ds \right) \end{pmatrix} \quad (20b)$$

To make the analysis more compact, we introduce the operators  $\mathcal{L}_{\sigma,K}^1$  and  $\mathcal{L}_{\sigma,K}^2$  that are defined on the set of  $N_K$  degrees of freedom in  $K$ , namely

$$\mathcal{L}_{\sigma,K}^1(V_1, \dots, V_r) = \begin{pmatrix} |C_{\sigma,K}|(V_r^\sigma - V_{r-1}^\sigma) + \int_{t_{n,0}}^{t_{n,r}} \mathcal{I}_0(\Phi_\sigma^x(V), s) ds \\ \vdots \\ |C_{\sigma,K}|(V_1^\sigma - V_0^\sigma) + \int_{t_{n,0}}^{t_{n,1}} \mathcal{I}_0(\Phi_\sigma^x(V), s) ds \end{pmatrix}, \quad (21a)$$

and

$$\mathcal{L}_{\sigma,K}^2(V_1, \dots, V_r) = \begin{pmatrix} \int_K \Psi_\sigma(V_r - u) dx + \int_{t_{n,0}}^{t_{n,r}} \mathcal{I}_{r+1}(\Phi_\sigma^x(V), s) ds \\ \vdots \\ \int_K \Psi_\sigma(V_1 - u) dx + \int_{t_{n,0}}^{t_{n,1}} \mathcal{I}_{r+1}(\Phi_\sigma^x(V), s) ds \end{pmatrix} \quad (21b)$$

Clearly we have  $\mathcal{L}_\sigma^1 = \sum_K \mathcal{L}_{\sigma,K}^1$  if

$$|C_\sigma| = \sum_{\sigma \ni K} |C_{\sigma,K}|$$

and we also have  $\mathcal{L}_\sigma^2 = \sum_K \mathcal{L}_{\sigma,K}^2$ .

Last, we define the operators  $\mathcal{L}^1$  and  $\mathcal{L}^2$  defined on  $V_h$ , the finite element set where the solution is sought for, as

$$\mathcal{L}^1 = (\mathcal{L}_\sigma^1)_{\sigma \in \mathcal{S}}, \quad \mathcal{L}^2 = (\mathcal{L}_\sigma^2)_{\sigma \in \mathcal{S}}.$$

We also introduce space and time operators for  $\mathcal{L}_{\sigma,K}^1$  and  $\mathcal{L}_{\sigma,K}^2$  defined component by component:

$$(\mathcal{L}_{\sigma,K,t}^1)_p = |C_{\sigma,K}|(V_{p+1} - V_p), \quad (\mathcal{L}_{\sigma,K,t}^2)_p = \int_K \Psi_\sigma(V_p - V_0) dx \quad (22)$$

and

$$(\mathcal{L}_{\sigma,K,\mathbf{x}}^1)_p = \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_0(\Phi_\sigma^\mathbf{x}(V, s)) ds, \quad (\mathcal{L}_{\sigma,K,\mathbf{x}}^2)_p = \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_{r+1}(\Phi_\sigma^\mathbf{x}(V, s)) ds. \quad (23)$$

More explicitly, we have as in the ODE case, expressions for each step:

- For the Euler step, the  $p$ -th component of  $\mathcal{L}^1$  is:

$$\mathcal{L}_\sigma^1(V_1, \dots, V_r)_p = |C_\sigma|(V_p^\sigma - V_0^\sigma) + \Delta t \sum_{K \ni \sigma} \left( \sum_{l=1}^{p-1} \xi_l \Phi_\sigma^\mathbf{x}(V_l) \right)$$

It is purely explicit and solving  $\mathcal{L}^1(V) = 0$  amounts to solving several Euler forward steps.

- For the corrector  $\mathcal{L}^2$ , the  $p$ -th component of  $\mathcal{L}^2$  is

$$\mathcal{L}_\sigma^2(V_1, \dots, V_{r+1})_p = \sum_{K \ni \sigma} \left( \int_K \Psi_\sigma(V_p - V_0) dx + \sum_{l=0}^{r+1} \theta_{l,r} \mathcal{I}_{r+1}(\Phi_\sigma^\mathbf{x}(V_l)) \right).$$

Then, we evaluate  $u_\sigma^{n+1}$  as in the ODE case:

- Evaluate  $V^{(0)} = (V_1^{(0)}, \dots, V_{r+1}^{(0)})$  as the solution of  $\mathcal{L}_\sigma^1(V^{(0)}) = 0$ . This amounts to using the Euler forward method.
- Knowing  $V^{(m)} = (V_1^{(m)}, \dots, V_{r+1}^{(m)})$ ,  $m > 0$ , evaluate  $V^{m+1} = (V_0^{(m+1)}, \dots, V_{r+1}^{(m+1)})$  as the solution of

$$\mathcal{L}_\sigma^1(V^{(m+1)}) = \mathcal{L}_\sigma^1(V^{(m)}) - \mathcal{L}_\sigma^2(V^{(m)}).$$

More explicitly, we have:

- Euler step: for  $p = 1, \dots, r$ , knowing that  $V_0 = u_\sigma^n$ ,

$$V_p^{(1)} = V_p^{(0)} - \frac{\Delta t}{|C_\sigma|} \sum_{K \ni \sigma} \sum_{l=1}^{p-1} \alpha_p \Phi_\sigma^\mathbf{x}(V_l^{(1)}).$$

- Correction step  $\#m$ :

$$\begin{aligned} & |C_\sigma|(V_p^{(m+1)} - V_0) + \Delta t \sum_{K \ni \sigma} \sum_{l=1}^{p-1} \alpha_l \Phi_\sigma^\mathbf{x}(V_l^{(m+1)}) \\ &= |C_\sigma|(V_p^m - V_0) + \Delta t \sum_{K \ni \sigma} \sum_{l=1}^{p-1} \alpha_l \Phi_\sigma^\mathbf{x}(V_p^{(m)}) \\ & \quad - \sum_{K \ni \sigma} \left( \int_K \Psi_\sigma(V_p^{(m)} - V_0) dx - \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_{r+1}(\Phi_\sigma^\mathbf{x}(V^{(m)}; s)) \right) \end{aligned}$$

i.e.

$$\begin{aligned}
& |C_\sigma|(V_p^{(m+1)} - V_0) + \Delta t \sum_{K \ni \sigma} \sum_{l=1}^{p-1} \alpha_l \Phi_\sigma^{\mathbf{x}}(V_l^{(m+1)}) \\
&= |C_\sigma|(V_p^{(m)} - V_0) + \Delta t \sum_{K \ni \sigma} \sum_{l=1}^{p-1} \alpha_l \Phi_\sigma^{\mathbf{x}}(V_p^{(m)}) \\
&\quad - \sum_{K \ni \sigma} \left( \int_K \Psi_\sigma(V_p^{(m)} - V_0) dx - \sum_{l=1}^{r+1} \theta_{l,r+1} \mathcal{I}_{r+1}(\Phi_\sigma^{\mathbf{x}}(V^{(m)}); s) \right)
\end{aligned}$$

i.e. after simplifications,

$$V_p^{(m+1)} = V_p^{(m)} + \frac{\Delta t}{|C_\sigma|} \Phi^{p+1,m+1,m} \quad (24a)$$

with

$$\begin{aligned}
\Phi^{p+1,m+1,m} &= \sum_{l=1}^{p-1} \alpha_l \int \psi_\sigma \operatorname{div} \left( \mathbf{f}(V_l^{(m+1)}) - \mathbf{f}(V_l^{(m)}) \right) dx \\
&\quad - \left( \int \psi_\sigma (V_p^{(m)} - V_0) dx + \int \psi_\sigma \operatorname{div} \left( \sum_{l=1}^{r+1} \theta_{l,r+1} \mathbf{f}(V_l^{(m)}) \right) \right)
\end{aligned} \quad (24b)$$

In the case we use  $\mathcal{L}_\sigma^1$  defined by

$$\mathcal{L}_\sigma^1(V_1, \dots, V_{r+1}) = \begin{pmatrix} |C_\sigma|(V_{r+1} - V_0) + \Delta t \Phi_\sigma^{\mathbf{x}}(V_0) \\ |C_\sigma|(V_r - V_0) + \xi_r \Delta t \Phi_\sigma^{\mathbf{x}}(V_0) \\ \vdots \\ |C_\sigma|(V_1 - V_0) + \xi_1 \Delta t \Phi_\sigma^{\mathbf{x}}(V_0) \end{pmatrix} \quad (25)$$

then (24b), because  $\xi_p = \alpha_1 + \dots + \alpha_p$ , is modified into

$$\begin{aligned}
\Phi^{p+1,m+1,m} &= \sum_{l=1}^p \alpha_l \int \psi_\sigma \operatorname{div} \left( \mathbf{f}(V_l^{(m+1)}) - \mathbf{f}(V_0) \right) dx \\
\Phi^{p+1,m+1,m} &= - \int \psi_\sigma (V_p^{(m)} - V_0) dx - \int \psi_\sigma \operatorname{div} \left( \sum_{l=1}^{r+1} \theta_{l,r+1} \mathbf{f}(V_l^{(m)}) \right)
\end{aligned} \quad (26)$$

The advantage of (24a)-(26) with respect to (24a)-(24b) is that one can compute only one residual for the Euler step.

For now, we have worked in a very formal way, for example several assumptions were implicitly done for all this to be defined. In the next section, we give a more precise statement on the assumptions we make, and also answer the main question: how to choose  $|C_\sigma|$  and to check what are the conditions on the finite element space to get a meaningful approximation.

## 4.2 Analysis

### 4.2.1 Introduction

It is well known that

**Proposition 4.2.** *If two operators  $\mathcal{L}_\Delta^1$  and  $\mathcal{L}_\Delta^2$  defined on  $\mathbb{R}^m$ , which depend of a parameter  $\Delta$ , are such that:*

1. *There exists  $\alpha_1 > 0$  independent of  $\Delta$  such that for any  $U, V$ ,*

$$\alpha_1 \|U - V\| \leq \|\mathcal{L}_\Delta^1(U) - \mathcal{L}_\Delta^1(V)\|, \quad (27)$$

2. *There exists  $\alpha_2 > 0$  independent of  $\Delta$  such that for any  $U, V$ ,*

$$\left\| (\mathcal{L}_\Delta^1(U) - \mathcal{L}_\Delta^2(U)) - (\mathcal{L}_\Delta^1(V) - \mathcal{L}_\Delta^2(V)) \right\| \leq \alpha_2 \Delta \|U - V\|. \quad (28)$$

*This last condition is nothing more than saying that the operator  $\mathcal{L}_\Delta^1 - \mathcal{L}_\Delta^2$  is uniformly Lipschitz continuous with Lipschitz constant  $\alpha_2 \Delta$*

*We also assume there exist a unique  $U_\Delta^*$  such that  $\mathcal{L}_\Delta^2(U_\Delta^*) = 0$ . Then if  $\nu = \alpha_2 \Delta < 1$  the defect correction is convergent, and after  $p$  iterations the error is smaller than  $\frac{\nu^p}{\alpha_1}$ .*

We recall the proof for the sake of completeness.

*Proof.* We drop the dependency in term of  $\Delta$  to simplify the text. Let us denote by  $U^*$  the solution of  $\mathcal{L}^2(U^*) = 0$ . We obviously have  $\mathcal{L}^1(U^*) = \mathcal{L}^1(U^*) - \mathcal{L}^2(U^*)$ , so that

$$\begin{aligned} \mathcal{L}^1(U^{m+1}) - \mathcal{L}^1(U^*) &= (\mathcal{L}^1(U^m) - \mathcal{L}^1(U^*)) - (\mathcal{L}^2(U^m) - \mathcal{L}^2(U^*)) \\ &= (\mathcal{L}^1(U^m) - \mathcal{L}^2(U^m)) - (\mathcal{L}^1(U^*) - \mathcal{L}^2(U^*)) \end{aligned}$$

so that

$$\begin{aligned} \alpha \|U^{m+1} - U^*\| &\leq \|\mathcal{L}^1(U^{m+1}) - \mathcal{L}^1(U^*)\| = \|(\mathcal{L}^1(U^m) - \mathcal{L}^2(U^*)) - (\mathcal{L}^1(U^*) - \mathcal{L}^2(U^*))\| \\ &\leq \|\mathcal{L}^1 - \mathcal{L}^2\| \|U^m - U^*\| \\ &\leq \alpha_2 \Delta \|U^m - U^*\| \leq (C\Delta)^m \|U^0 - U^*\| \end{aligned}$$

So we see that after  $m$  iteration, we have an error at most  $(\alpha_2 \Delta)^m / \alpha_1$ . □

Let us turn back to our problem. The question is twofold:

- Under which condition, the  $\mathcal{L}^1$  operator is invertible and satisfy an inequality of the type (27) ?
- Under which conditions do we have

$$\|\mathcal{L}^1 - \mathcal{L}^2\| = O(\Delta t) + O(h). \quad (29)$$

Here the parameter  $\Delta$  is  $h + \Delta t$ .

### 4.2.2 Choice of norms

First, we need to define a relevant norm. We equip  $V^h$  with the  $H^1$  norm:

$$\|v\|_{H^1(\Omega)}^2 = \int_{\Omega} v^2 d\mathbf{x} + \int_{\Omega} \nabla v^2 d\mathbf{x},$$

and, for any two  $u \in \mathcal{L}^2(\Omega)$ ,  $\Phi \in \mathcal{L}^2(\Omega)$ , we set

$$\langle u, \Phi \rangle = \int_{\Omega} u(\mathbf{x})\Phi(\mathbf{x})d\mathbf{x}.$$

It is well known from the Cauchy-Schwarz inequality and the fact that  $H^1(\Omega) \subset L^2(\Omega)$  that the quantity

$$\|\Phi\| = \sup_{u \in H^1(\Omega)} \frac{\langle u, \Phi \rangle}{\|u\|_{H^1(\Omega)}} \quad (30a)$$

defines a norm on  $L^2(\Omega)$ . For  $\Phi = (\Phi_1, \dots, \Phi_m) \in (L^2(\Omega))^m$ , we set

$$\|\Phi\| = \max_{i=1, \dots, m} \|\Phi_m\|. \quad (30b)$$

### 4.2.3 Galerkin scheme

It is useful in the following to introduce the Galerkin residuals:

$$\Phi(V_{l,K})_{\sigma}^{G,\mathbf{x}} = - \int_K \varphi_{\sigma} \mathbf{f}(V_{l,K}) d\mathbf{x} + \int_{\partial K} \varphi_{\sigma} \mathbf{f}(V_{l,K}) \cdot \mathbf{n}. \quad (31)$$

We see that

$$\sum_{\sigma \in K} \Phi(V_{l,K})_{\sigma}^{G,\mathbf{x}} = \int_K \mathbf{f}(V_{l,K}) \cdot \mathbf{n} = \sum_{\sigma \in K} \Phi(V_{l,K})_{\sigma}^{\mathbf{x}} \quad (32)$$

for any of the spatial residuals defined above.

### 4.2.4 Coercivity of $\mathcal{L}^1$

We have a first result on the behavior of  $\mathcal{L}^1$ .

**Lemma 4.3.** *We assume that the residuals  $\Phi_{\sigma}^{\mathbf{x}}$  are Lipschitz continuous.*

1. *If  $|C_{\sigma}| > 0$  for any  $\sigma \in \Sigma$ , then  $\mathcal{L}^1$  is invertible.*
2. *From now on, we assume in addition:*

$$\begin{aligned} |C_{\sigma,K}| &> 0 && \text{if } \sigma \in K, \\ |C_{\sigma,K}| &= 0 && \text{else.} \end{aligned} \quad (33a)$$

*We also assume*

$$|C_{\sigma}| = \sum_{K \ni \sigma} |C_{\sigma,K}|. \quad (33b)$$

If there exists constants  $C_1 > 0$  and  $C_2 > 0$ , independent of the mesh family such that

$$C_1 \leq \frac{|C_{\sigma,K}|}{|K|} \leq C_2, \quad (34)$$

then there exists  $\alpha_1 > 0$  independent of the mesh family such that for any  $U$  and  $V$  in (20) and (21).

$$\alpha \|U - V\| \leq \|\mathcal{L}^1(U) - \mathcal{L}^1(V)\|. \quad (35)$$

Before proving this lemma, we have a first one:

**Lemma 4.4.** *Under the conditions of lemma 4.3, the solution of  $\mathcal{L}_\sigma^1(V_1, \dots, V_{r+1}) = |C_\sigma|(A_1, \dots, A_{r+1})^T$  satisfies*

$$\begin{aligned} V_1 &= A_1 + V_0, \\ V_2 &= H_2(V_0 + A_1, A_2), \\ &\vdots \\ V_l &= H_l(V_0 + A_1, A_2, \dots, A_l), \\ &\vdots \\ V_{r+1} &= H_{r+1}(V_0 + A_1, A_2, \dots, A_{r+1}) \end{aligned} \quad (36)$$

where the  $H_j$  are Lipschitz continuous with respect to their arguments.

*Proof.* The proof is immediate thanks to the explicit nature of the scheme and condition (34), and because the residuals  $\Phi_\sigma^x$  are Lipschitz continuous.  $\square$

*Proof of lemma 4.3.* 1. Clearly, if  $A = (A^\sigma)_{\sigma \in \Sigma}$  is a vector of  $\mathbb{R}^{|\Sigma|}$ , we can solve  $\mathcal{L}^1(V) = A$  if and only if  $|C_\sigma| \neq 0$ . Since  $C_\sigma > 0$  is positive, we have our first condition.

2. Consider  $U$  and  $V$  in  $\mathbb{R}^{|\Sigma|}$ , and we are interested in

$$\|\mathcal{L}^1(U) - \mathcal{L}^1(V)\| = \max_{v \in H^1(\Omega)} \frac{\sum_\sigma v_\sigma (\mathcal{L}_\sigma^1(U) - \mathcal{L}_\sigma^1(V))}{\|v\|_{H^1}}.$$

We set  $\mathcal{L}^1(U) = |C_\sigma|A$  and  $\mathcal{L}^1(V) = |C_\sigma|B$  where  $A = (A_l^\sigma)_{\sigma \in \Sigma, 1 \leq l \leq r+1}$  and  $B = (B_l^\sigma)_{\sigma \in \Sigma, 1 \leq l \leq r+1}$ .

In order to prove (35), it is sufficient to show a similar condition on each of the components of  $\mathcal{L}^1$ . We first have:

$$\begin{aligned} |C_\sigma|(U_p^\sigma - U_0^\sigma) &= |C_\sigma|A_p - \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_0(\Phi_\sigma^x(U_0), \Phi_\sigma^x(U_1), \dots, \Phi_\sigma^x(U_{p-1})) \\ |C_\sigma|(V_p^\sigma - V_0^\sigma) &= |C_\sigma|B_p - \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_0(\Phi_\sigma^x(V_0), \Phi_\sigma^x(V_1), \dots, \Phi_\sigma^x(V_{p-1})) \end{aligned}$$

and, because  $\mathcal{I}_0$  is linear,

$$|C_\sigma|(U_p^\sigma - V_p^\sigma) = |C_\sigma|(A_p - B_p) - \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_0(0, \Phi_\sigma^x(U_1) - \Phi_\sigma^x(V_1), \dots, U_{p-1} - \Phi_\sigma^x(V_{p-1})).$$

We introduce the notation:

$$\Phi_\sigma^{\mathbf{x}}(U_p) - \Phi_\sigma^{\mathbf{x}}(V_p) := \delta_p \Phi_{\sigma,K}^{\mathbf{x}}.$$

Then we multiply by a test function, and get

$$\begin{aligned} \sum_{\sigma} v_{\sigma} |C_{\sigma}| (U_p^{\sigma} - V_p^{\sigma}) &= \sum_{\sigma} v_{\sigma} |C_{\sigma}| (A_p^{\sigma} - B_p^{\sigma}) - \sum_{\sigma} v_{\sigma} \sum_{K \ni \sigma} \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_0(\delta_1 \Phi_{\sigma,K}^{\mathbf{x}}, \dots, \delta_p \Phi_{\sigma,K}^{\mathbf{x}}) \\ &= \sum_{\sigma} v_{\sigma} |C_{\sigma}| (A_p^{\sigma} - B_p^{\sigma}) - \sum_K \sum_{K \ni \sigma} v_{\sigma} \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_0(\delta_1 \Phi_{\sigma,K}^{\mathbf{x}}, \dots, \delta_p \Phi_{\sigma,K}^{\mathbf{x}}) \end{aligned}$$

Using the Galerkin residuals and the notation  $\Delta \Phi_{\sigma}(U) := \Phi_{\sigma}(U) - \Phi_{\sigma}^{Gal}(U)$ , we obtain

$$\begin{aligned} \sum_{K \ni \sigma} v_{\sigma} \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_0(\delta_1 \Phi_{\sigma}^{\mathbf{x}}, \dots, \delta_p \Phi_{\sigma}^{\mathbf{x}}) &= \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_0 \left( \int_K \nabla v \cdot (\mathbf{f}(U_1) - \mathbf{f}(V_1)) d\mathbf{x}, \dots, \right. \\ &\quad \left. \int_K \nabla v \cdot (\mathbf{f}(U_p) - \mathbf{f}(V_p)) d\mathbf{x} \right) d\mathbf{x} \\ &\quad + \sum_{\sigma} (v_{\sigma} - v_{\sigma'}) \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_0(\Delta \Phi_{\sigma}^{\mathbf{x}}(U_1) - \Delta \Phi_{\sigma}^{\mathbf{x}}(V_1), \dots, \Delta \Phi_{\sigma}^{\mathbf{x}}(U_{p-1}) - \Delta \Phi_{\sigma}^{\mathbf{x}}(V_{p-1})) \end{aligned}$$

Here  $\sigma'$  is any fixed degree of freedom in  $K$ . Since the flux  $\mathbf{f}$  is Lipschitz continuous, as well as the residuals, and since we see that:

$$\begin{aligned} \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_0 \left( \int_K \nabla v \cdot (\mathbf{f}(U_1) - \mathbf{f}(V_1)) d\mathbf{x}, \dots, \int_K \nabla v \cdot (\mathbf{f}(U_p) - \mathbf{f}(V_p)) d\mathbf{x} \right) d\mathbf{x} \\ = \sum_{l=1}^p \alpha_l \int_K \nabla v \cdot (\mathbf{f}(U_l) - \mathbf{f}(V_l)) d\mathbf{x}, \end{aligned}$$

we have

$$\begin{aligned} \left| \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_0 \left( \int_K \nabla v \cdot (\mathbf{f}(U_1) - \mathbf{f}(V_1)) d\mathbf{x}, \dots, \int_K \nabla v \cdot (\mathbf{f}(U_p) - \mathbf{f}(V_p)) d\mathbf{x} \right) d\mathbf{x} \right| \\ \leq \|v\|_{H^1} \sum_{l=1}^p \alpha_l L \|U_l - V_l\|_2 \end{aligned}$$

Similarly, using lemma A.1,  $|\Phi_{\sigma,K}(U) - \Phi_{\sigma,K}(V)| \leq Lh_K \sum_{\sigma' \in K} |U_{\sigma} - V_{\sigma'}|$ , and then

$$\begin{aligned} \left| \sum_{\sigma} (v_{\sigma} - v_{\sigma'}) \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_0(\Delta \Phi_{\sigma}^{\mathbf{x}}(U_1) - \Delta \Phi_{\sigma}^{\mathbf{x}}(V_1), \dots, \Delta \Phi_{\sigma}^{\mathbf{x}}(U_{p-1}) - \Delta \Phi_{\sigma}^{\mathbf{x}}(V_{p-1})) \right| \\ \leq \sum_{l=1}^{p-1} \alpha_l \sum_{\sigma} |v_{\sigma} - v_{\sigma'}| |\Delta \Phi_{\sigma}^{\mathbf{x}}(U_l) - \Delta \Phi_{\sigma}^{\mathbf{x}}(V_l)| \\ \leq C_K \|\nabla v\|_{2,K} \sum_{l=1}^{p-1} \|U_l - V_l\|_{2,K} \end{aligned}$$

where  $C_K$  depends on the number of vertices of  $K$ . Then we conclude by using lemma 4.4 which states that  $|U_l - V_l|$  is bounded by  $C \sum_{j=1}^l |C_{\sigma}| |A_j - B_j|$ . Thanks to condition (34), we get the result.  $\square$

#### 4.2.5 Error estimate for $\mathcal{L}^1 - \mathcal{L}^2$

We write, for any  $\sigma$ ,  $\mathcal{L}_\sigma^\ell = (\mathcal{L}_{\sigma,0}^\ell, \mathcal{L}_{\sigma,1}^\ell, \dots, \mathcal{L}_{\sigma,m}^\ell)^T$  and look for:  $\max_{k=0,m} \|\mathcal{L}_k^1 - \mathcal{L}_k^2\|$ .

We have

$$\|\mathcal{L}_k^1(V) - \mathcal{L}_k^2(V)\| = \sup_{v \in H^1(\Omega)} \frac{\sum_\sigma v_\sigma (\mathcal{L}_\sigma^1(V) - \mathcal{L}_\sigma^2(V))}{\|v_h\|_{H^1}}.$$

Here,  $V = (V_1, \dots, V_{r+1})$  to simplify the notations. Since

$$\sum_\sigma v_\sigma (\mathcal{L}_\sigma^1(V) - \mathcal{L}_\sigma^2(V)) = \sum_K \sum_{\sigma \in K} v_\sigma (\mathcal{L}_\sigma^1(V) - \mathcal{L}_\sigma^2(V))$$

it is enough to look at  $\sum_{\sigma \in K} v_\sigma (\mathcal{L}_\sigma^1(V) - \mathcal{L}_\sigma^2(V))$ .

We can write

$$\mathcal{L}_{\sigma,p}^1(V) = |C_{\sigma,K}|(U_\sigma^m - u_\sigma^n) + \int_{t_n}^{t_{n,p}} \int_K \psi_\sigma \mathcal{I}_0(\Phi_\sigma^{\mathbf{x}}(V_0), \dots, \Phi_\sigma^{\mathbf{x}}(V_{r+1})) ds,$$

$$\mathcal{L}_{\sigma,p}^2(V) = \int_K \psi_\sigma (V_p - V_0) + \int_{t_n}^{t_{n,p}} \mathcal{I}_l(\Phi_\sigma^{\mathbf{x}}(V_0), \dots, \Phi_\sigma^{\mathbf{x}}(V_{r+1})) ds$$

so that

$$\begin{aligned} \mathcal{L}_{\sigma,p}^1(V) - \mathcal{L}_{\sigma,p}^2(V) &= |C_{\sigma,K}|(U_\sigma^m - u_\sigma^n) - \int_K \psi_\sigma (U^m - u^n) \\ &\quad + \int_{t_n}^{t_{n,p}} \left( \mathcal{I}_0(\Phi_\sigma^{\mathbf{x}}(V_0), \dots, \Phi_\sigma^{\mathbf{x}}(V_{r+1})) - \mathcal{I}_l(\Phi_\sigma^{\mathbf{x}}(V_0), \dots, \Phi_\sigma^{\mathbf{x}}(V_{r+1})) \right) ds \end{aligned}$$

We see that to get the estimate (29), a sufficient condition is:

$$\sum_{\sigma \in K} |C_{\sigma,K}|(V_m^\sigma - V_0^\sigma) = \int_K (V_m - V_0) d\mathbf{x} \quad (37)$$

so that:

$$\sum_{\sigma \in K} |C_{\sigma,K}|(V_m^\sigma - V_0^\sigma) = \sum_{\sigma \in K} \int_K \psi_\sigma (V_m - V_0)$$

because  $\sum_{\sigma \in K} \psi_\sigma = 1 = \sum_{\sigma \in K} \varphi_\sigma$

**Proposition 4.5.** *Under the assumptions of lemma 4.3, there exists  $C > 0$  such that*

$$\|\mathcal{L}^1(V) - \mathcal{L}^2(V)\| \leq C(h + \Delta t) \|V\|.$$

*Proof.* The proof is rather similar to that of lemma 4.3. We first have

$$\begin{aligned} \mathcal{L}_{\sigma,p}^1(V) - \mathcal{L}_{\sigma,p}^2(V) &= |C_{\sigma,K}|(U_\sigma^m - u_\sigma^n) - \int_K \psi_\sigma (U^m - u^n) \\ &\quad + \int_{t_n}^{t_{n,p}} \left( \mathcal{I}_0(\Phi_\sigma^{\mathbf{x}}(V_0), \dots, \Phi_\sigma^{\mathbf{x}}(V_{r+1})) - \mathcal{I}_l(\Phi_\sigma^{\mathbf{x}}(V_0), \dots, \Phi_\sigma^{\mathbf{x}}(V_{r+1})) \right) ds \end{aligned}$$



so that

$$\begin{aligned} \sum_{\sigma \in K} v_\sigma (\mathcal{L}_{\sigma,p}^1(V) - \mathcal{L}_{\sigma,p}^2(V)) &= |C_{\sigma,K}| \sum_{\sigma \in K} v_\sigma (V_m^\sigma - V_0^\sigma) - \int_K \psi_\sigma (V_m - V_0) \\ &\quad + \sum_{\sigma \in K} \int_{t_n}^{t_{n+1}} \left( \mathcal{I}_0(\Phi_\sigma^{\mathbf{x}}(V_0), \dots, \Phi_\sigma^{\mathbf{x}}(V_{r+1})) - \mathcal{I}_l(\Phi_\sigma^{\mathbf{x}}(V_0), \dots, \Phi_\sigma^{\mathbf{x}}(V_{r+1})) \right) ds \end{aligned}$$

Let us have a look at  $\sum_{\sigma \in K} v_\sigma (V_m^\sigma - V_0^\sigma) - \int_K \psi_\sigma (V_m - V_0)$ . we have, for any  $\sigma_0 \in K$ :

$$\begin{aligned} \sum_{\sigma \in K} v_\sigma (V_m^\sigma - V_0^\sigma) - \int_K \psi_\sigma (V_m - V_0) &= v_{\sigma_0} \sum_{\sigma \in K} (V_m^\sigma - V_0^\sigma) - \int_K \psi_\sigma (V_m - V_0) \\ &\quad + \sum_{\sigma \in K} (v_\sigma - v_{\sigma_0}) (V_m^\sigma - V_0^\sigma) - \int_K \psi_\sigma (V_m - V_0) \\ &= \sum_{\sigma \in K} (v_\sigma - v_{\sigma_0}) (V_m^\sigma - V_0^\sigma) - \int_K \psi_\sigma (V_m - V_0) \end{aligned}$$

so that

$$\begin{aligned} \sum_K |C_{\sigma,K}| \left| \sum_{\sigma \in K} (v_\sigma - v_{\sigma_0}) (V_m^\sigma - V_0^\sigma) - \int_K \psi_\sigma (V_m - V_0) \right| &\leq h \left( \int_\Omega \|\nabla v\|^2 d\mathbf{x} \right)^{1/2} \left( \sum_K |C_{\sigma,K}| \sum_K (V_m^\sigma - V_0^\sigma)^2 \right)^{1/2} \\ &= Ch \|v\|_{H^1} \|V_m - V_0\|_2 \end{aligned}$$

Using condition (34) and lemma A.1.

The second term is handled similarly, with the same technique as in the proof of lemma 4.3:

$$\begin{aligned} \sum_{\sigma \in K} v_\sigma \int_{t_n}^{t_{n+1}} \left( \mathcal{I}_0(\Phi_\sigma^{\mathbf{x}}(V_0), \dots, \Phi_\sigma^{\mathbf{x}}(V_{r+1})) - \mathcal{I}_l(\Phi_\sigma^{\mathbf{x}}(V_0), \dots, \Phi_\sigma^{\mathbf{x}}(V_{r+1})) \right) ds \\ = \sum_{\sigma \in K} v_\sigma \int_{t_n}^{t_{n+1}} \left( \mathcal{I}_0(\nabla v \cdot \mathbf{f}(V_0), \dots, \nabla v \cdot \mathbf{f}(V_{r+1})) - \mathcal{I}_l(\Phi_\sigma^{G,\mathbf{x}}(V_0), \dots, \Phi_\sigma^{G,\mathbf{x}}(V_{r+1})) \right) \\ + \sum_{\sigma \in K} \int_{t_n}^{t_{n+1}} (v_\sigma - v_{\sigma'}) \left( \mathcal{I}_0(\Delta \Phi_\sigma^{\mathbf{x}}(V_0), \dots, \Delta \Phi_\sigma^{G,\mathbf{x}}(V_{r+1})) - \mathcal{I}_l(\Delta \Phi_\sigma^{\mathbf{x}}(V_0), \dots, \Delta \Phi_\sigma^{\mathbf{x}}(V_{r+1})) \right) \end{aligned}$$

Since clearly for any  $U_j$ ,

$$\|\mathcal{I}_0(U_0, \dots, U_{r+1}) - \mathcal{I}_l(U_0, \dots, U_{r+1})\|_2 \leq C \Delta t \sum_{l=1}^{r+1} \|U_l\|_2,$$

since the flux  $\mathbf{f}$  and the residuals are Lipschitz continuous, we get the result using again the same estimates as in the proof of lemma 4.3.  $\square$

As a consequence, we see that not all finite element can be used with this technique: A sufficient condition is that

$$\int_K \varphi_\sigma dx > 0 \tag{38}$$

This condition is met for any  $\mathbb{Q}_r$  approximation where the degree of freedom correspond to Gaussian points, for example a Cartesian product of one dimensional Gaussian points. For simplicies, we know that the integral of Lagrange basis functions can be of both sign and even vanish: think of the quadratic case for a triangle. This is why we consider, as in [6, 10] Bézier approximations: if  $\lambda_1, \dots, \lambda_{d+1}$  are the barycentric coordinates with respect to the vertices of a simplex, we define, for the multi-index  $(i_1, \dots, i_{d+1})$  with  $i_1 + \dots + i_{d+1} = r$

$$B_{i_1, \dots, i_{d+1}}(\mathbf{x}) = \frac{r!}{i_1! \dots i_{d+1}!} \lambda_1^{i_1}(\mathbf{x}) \dots \lambda_{d+1}^{i_{d+1}}(\mathbf{x}).$$

We have

$$\sum_{i_1, \dots, i_d, \sum i_j = r} B_{i_1, \dots, i_{d+1}}(\mathbf{x}) = 1$$

because  $\sum_{j=1}^{d+1} \lambda_j = 1$ . In addition for  $\mathbf{x} \in K$ ,  $B_{i_1, \dots, i_{d+1}}(\mathbf{x}) \geq 0$  and

$$\int_K B_{i_1, \dots, i_{d+1}}(\mathbf{x}) d\mathbf{x} > 0.$$

Similar properties can be stated for NURBS. In the case of Bézier, the  $B_{i_1, \dots, i_{d+1}}$  span  $\mathbb{P}^r(K)$ , but in order to indicate clearly which type of approximation we use, we denote by  $\mathbb{B}_r$  the space  $\mathbb{P}_r$  when it is spanned by Bézier polynomials.

### 4.3 Stability restriction on the time step.

For each iteration, the scheme is written as

$$\mathcal{L}^1(V^{(m+1)}) = \mathcal{L}^1(V^{(m)}) - \mathcal{L}^2(V^{(m)}) = O(h)$$

so that after  $K$  iterations, we get:

$$\mathcal{L}^1(V^{(K)}) = O(h).$$

Since  $\mathcal{L}^1(V) = 0$  essentially amounts to a two level scheme for each of the sub-time steps  $t_{n,m}$ ,  $m = 0, \dots, K$ , we see that the solution  $V_k^K$  is obtained from a two-level schemes that is perturbed by an  $O(h)$  term. From a result in [20], we see that, give a norm, the stability condition of the method is that of  $\mathcal{L}^1$ . Since the Euler forward method is used, we see that for a method of order  $r$  in space, the time step must be divided by  $r$  with respect to the time step needed for the first order in space scheme.

### 4.4 Maximum principle

Here we are interested in the maximum principle. We drop the time subscript,  $l$  is the sub-time subscript,  $u_\sigma$  is the solution at the beginning of the computation. In what follows, we use  $\mathcal{L}^1$  defined by (25). The  $\mathcal{L}^2$  operator, i.e. for each element and each subtime step  $p$ ,

$$\int_K \psi_\sigma (V_p^{(l)} - V_0) dx + \int_{\partial K} \psi_\sigma \left( \sum_{k=1}^{r+1} \theta_{k,r+1} \mathbf{f}(V_k^{(l)}) \right) \cdot \mathbf{n}$$

writes

$$|K| \sum_{\sigma' \in K} a_{\sigma\sigma'}^K (V_{p,\sigma'}^{(l)} - u_{\sigma'}) + \Delta t \sum_{k=1}^{r+1} \theta_{k,r+1} \left\{ \sum_{\sigma' \in K} c_{\sigma\sigma'}^k (V_{k,\sigma}^{(l)} - V_{k,\sigma'}^{(l)}) \right\}.$$

Here we assume that

- $a_{\sigma\sigma}^K = \gamma_\sigma^K \frac{|K|}{\#DOF}$  with  $\gamma_\sigma^K \in [0, 1]$  and  $\#DOF$  is the number of degrees of freedom in  $K$
- and  $c_{\sigma\sigma'}^k \geq 0$ .

This can be made possible by using the nonlinear RDS schemes described in sections 4 and 5.3, see annex B. We can also assume that  $\theta_{k,r+1} \geq 0$ , for example by using the strong stability preserving Deferred correction schemes of [17]; they have been designed up to fourth order accuracy. The iterative steps are then:

$$\begin{aligned} |C_\sigma| (V_{p,\sigma}^{l+1} - u_\sigma) &= |C_\sigma| (V_{p,\sigma}^l - u_\sigma) \\ &\quad - \left( \sum_{k \ni \sigma} |K| \gamma_\sigma^K \frac{|K|}{\#DOF} (V_{p,\sigma}^{(l)} - u_\sigma) + \Delta t \sum_{k=1}^K \omega_k \left\{ \sum_{\sigma' \neq \sigma} c_{\sigma\sigma'}^k (V_{k,\sigma}^{(l)} - V_{k,\sigma'}^{(l)}) \right\} \right) \end{aligned} \quad (39)$$

with  $c_{\sigma\sigma'}^k \geq 0$ ,  $K$  the number of sub-time steps. This can be rewritten as:

$$\begin{aligned} |C_\sigma| V_{p,\sigma}^{l+1} &= \left( |C_\sigma| - \sum_{k \ni \sigma} |K| \gamma_\sigma^K \frac{|K|}{\#DOF} - \Delta t \sum_{k=1}^K \omega_k \left\{ \sum_{\sigma' \neq \sigma} c_{\sigma\sigma'}^k \right\} \right) V_{p,\sigma}^l \\ &\quad + \Delta t \sum_{k=1}^K \omega_k \sum_{\sigma' \neq \sigma} c_{\sigma\sigma'}^k V_{k,\sigma'}^{(l)} \\ &\quad + \sum_{k \ni \sigma} |K| \gamma_\sigma^K \frac{|K|}{\#DOF} u_\sigma. \end{aligned}$$

Since  $|C_\sigma| = \sum_{k \ni \sigma} \gamma_\sigma^K \frac{|K|}{\#DOF}$ ,  $\gamma_\sigma^K \in [0, 1]$  and if  $\omega_k \geq 0$ , we have a maximum principle under a CFL like condition.

## 5 Applications

We present results with the second, third and fourth order temporal schemes and  $\mathbb{P}_1$ ,  $\mathbb{B}_2$  and  $\mathbb{B}_3$  elements (the last one for the one dimensional for the one dimensional case). More specifically, the time schemes rely on the following quadrature formula:

- Second order in time: it relies on a linear Lagrange interpolation on  $[0, 1]$ , so:

$$\int_0^1 \mathcal{I}_1(f) ds = \frac{1}{2} (f(0) + f(1)).$$

This gives back the scheme of [19].

- Third order in time: It is based on the Lagrange interpolation in  $[0, 1]$ , where the data are given at the points  $t = 0, \frac{1}{2}$  and  $1$ . This results in the following formula that defines the operator  $\mathcal{L}^2$ :

$$\int_0^{1/2} \mathcal{I}_2(f) ds = \frac{5}{24} f(0) + \frac{1}{3} f\left(\frac{1}{2}\right) - \frac{1}{24} f(1)$$

$$\int_0^1 \mathcal{I}_2(f) ds = \frac{1}{6} f(0) + \frac{4}{6} f\left(\frac{1}{2}\right) + \frac{1}{6} f(1)$$

We have used the same temporal scheme for  $\mathbb{P}^1$  and  $\mathbb{B}^2$  elements.

- Fourth order in time: it relies on the Lagrange interpolation approximation with the points  $t = \frac{1}{2}(1 + \cos(k\pi/3))$ ,  $k = 0, \dots, 3$ . We have

$$\int_0^{1/4} \mathcal{I}_3(f) ds = \frac{59}{576} f(0) + \frac{47}{288} f\left(\frac{1}{4}\right) - \frac{7}{288} f\left(\frac{1}{2}\right) + \frac{5}{576} f(1)$$

$$\int_0^{3/4} \mathcal{I}_3(f) ds = \frac{3}{64} f(0) + \frac{15}{32} f\left(\frac{1}{4}\right) + \frac{9}{32} f\left(\frac{3}{4}\right) - \frac{3}{64} f(1)$$

$$\int_0^1 \mathcal{I}_3(f) ds = \frac{1}{18} f(0) + \frac{4}{9} f\left(\frac{1}{4}\right) + \frac{4}{9} f\left(\frac{3}{4}\right) + \frac{1}{18} f(1)$$

We also have used in experiments that are not reported the equidistributed sequence. Since the order is still low, it does not change the results.

The space and time residuals that are tested are:

- Burman's (16)
- SUPG (15)
- the nonlinear RDS schemes

$$\Phi_\sigma(u^h) = \beta_\sigma(u^h) \Phi_{tot} + \sum_{\text{edges of } \partial K} h_e^2 \|\nabla_u f\| \int_{\partial K} [\nabla u^h] \cdot [\nabla \varphi_\sigma] dl \quad (40)$$

where the choice of  $\beta_\sigma(u^h)$  is done as follows

- Choice 1:

$$\beta_\sigma^{PSI}(u^h) = \frac{\left(\frac{\Phi_\sigma^{LxF}}{\Phi_{tot}}\right)^+}{\sum_{\sigma' \in K} \left(\frac{\Phi_{\sigma'}^{LxF}}{\Phi_{tot}}\right)^+}$$

Here the scheme is named PSI (for historical reasons),

- or Choice 2:

$$\beta_\sigma(u^h) \Phi_{tot} = (1 - \theta) \beta_\sigma^{PSI}(u^h) \Phi_{tot} + \theta \Phi_\sigma^{LxF}$$

$$\theta = \frac{|\Phi_{tot}|}{\sum_{\sigma' \in K} |\Phi_{\sigma'}^{LxF}|}$$

here it is denoted as Blended PSI.

Other choices are possible, for example to blend the Burman or SUPG residual with a Lax Friedrichs one, the blending parameter is defined by  $\theta$  in choice 2. We nickname this schemes as Blended Burman with Jump stabilisation (or Blended SUPG with Jump stabilisation).

## 5.1 A one dimensional example.

The test case is the problem

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$$

on  $[0, 1]$  with Neuman boundary conditions. The initial condition is

$$u(x, 0) = e^{-80(x-0.4)^2}.$$

The results are given on table 1: we obtain the expected accuracy.

We also have considered the initial condition given by :

$$u_0(x) = \begin{cases} (G(x, \beta, z - \delta) + G(x, \beta, z + \delta) + 4G(x, \beta, z))/6. & \text{if } -0.8 \leq x \leq -0.6 \\ 1 & \text{if } -0.4 \leq x \leq -0.2 \\ 1 - |10(x - 1)| & \text{if } 0 \leq x \leq 0.2 \\ \frac{1}{6}(F(x, \alpha, a - \delta) + F(x, \alpha, a + \delta) + 4F(x, \alpha, z)) & \text{if } 0.4 \leq x \leq 0.6 \\ 0 & \text{otherwise} \end{cases} \quad (41)$$

with  $G(x, \beta, z) = e^{-\beta(x-z)^2}$  and  $F(x, \alpha, a) = \sqrt{\max(1 - \alpha^2(x - a)^2, 0)}$ . In the present case, we have taken  $a = 0.5$ ,  $z = -0.7$ ,  $\delta = 0.005$ ,  $\alpha = 10$  and  $\beta = \log(2)/(36\delta^2)$  as in [15, 14]. The boundary conditions are periodic, and we are looking at the solution at  $t = 8$ . This is quite a challenging case since the solution presents smooth parts and discontinuous features simultaneously.

We have first run the method for quadratic elements and the Galerkin+Jump scheme. The final time is  $T = 8$ . For 3 iterations the results are plagued by oscillations that are created by the discontinuous figures. Note that for perfectly regular solutions (as in the previous case), the scheme with 3 iteration is fine. For 4 iterations, the Gaussian bump is very well represented, but we still have oscillations where the solution should be vanishing. This has completely disappeared with 5 iterations and more. We note that the smooth parts of the solutions are always very well represented, and the resolution improves for the discontinuous part when we increase the number of iterations. Then we have run the same case with non linear schemes, the results are displayed in figure 2. Only 3 iterations are used, the final time is  $T = 8$ . The results are non oscillatory as expected, and of course more dissipated as in the Galerkin+Jump scheme. This is not a surprise.

## 5.2 2D linear example

The velocity field at  $(x, y)$  is given by  $\mathbf{a} = 2\pi(-y, x)$ . The initial condition is given by:

$$u_0(x, y) = e^{-40(x^2+y^2)}.$$

The domain is a circle with center  $(0, 0)$  and radius  $R = 1$ . The mesh representing all the degrees of freedom is displayed in Figure 3: the quadratic elements have 6 degrees of freedom (the vertices and the mid-points of the edges). These degrees of freedom are also used for the linear element just by mesh refinement. There are 7047 degrees of freedom here, so  $h \approx \sqrt{\frac{\pi}{7047}} \approx 0.021$  which is

$\mathbb{B}_2$			$\mathbb{B}_3$	
$L^1$ error				
$\log_{10} h$	$\log_{10}(err)$	slope	$\log_{10}(err)$	slope
-1.00000000	-1.77300358	-	-.2088D+01	-
-1.30103004	-3.09956264	4.40673304	-.4004D+01	6.35
-1.60205996	-3.92077780	2.72801828	-.5237D+01	4.09
-1.90309000	-4.74021816	2.72212148	-.6425D+01	4.09
-2.20411992	-5.61343765	2.90077305	-.7632D+01	4.09
-2.50515008	-6.50857687	2.97358656	-.8835D+01	4.09
$L^2$ error				
$\log_{10} h$	$\log_{10}(err)$	slope	$\log_{10}(err)$	slope
-1.00000000	-1.63971663	-	-.2310D+01	-
-1.30103004	-2.85121393	4.02450657	-.3498D+01	3.94
-1.60205996	-3.66986251	2.71949244	-.4681D+01	3.93
-1.90309000	-4.47560310	2.67661190	-.5879D+01	3.98
-2.20411992	-5.34178400	2.87739134	-.7082D+01	3.99
-2.50515008	-6.23404074	2.96401119	-.8286D+01	3.99
$L^\infty$ error				
$\log_{10} h$	$\log_{10}(err)$	slope	$\log_{10}(err)$	slope
-1.00000000	-1.23617887	-	-.1617D+01	-
-1.30103004	-2.30927277	3.56474018	-.2843D+01	4.07
-1.60205996	-3.08773041	2.58598089	-.3943D+01	3.65
-1.90309000	-3.87637067	2.61980581	-.5105D+01	3.86
-2.20411992	-4.73004484	2.83584499	-.6299D+01	3.96
-2.50515008	-5.61776066	2.94892645	-.7499D+01	3.96

Table 1: Errors for  $u(x, 0) = e^{-80(x-0.4)^2}$  and linear advection.  $t = 0.25$ .  $\mathbb{B}_2$ /3rd order in time,  $\mathbb{B}_3$ /4th order in time.

relatively coarse. On the same figure, we represent the exact solution. The time step is evaluated as the minimum of the  $\Delta t_K$  defined by:

$$\Delta t_K = CFL \frac{h_K}{\|\bar{\mathbf{a}}_K\|}$$

where  $h_K$  is the length of the smallest edge of  $K$  and  $\bar{\mathbf{a}}_K$  is the speed at the centroid. Since the elements for the  $\mathbb{P}^1$  simulations are obtained from those of the  $\mathbb{B}^2$  simulation by splitting, the parameter  $h_K$ , for the  $\mathbb{P}^1$  simulations, is half of the one for the  $\mathbb{B}^2$  simulations. For that reason, the CFL number for the quadratic approximation is half of the one chosen for the linear simulations, namely 0.6 instead of 0.3: we run with the same time step. By the way, we have not yet conducted a rigorous study of the CFL condition, but all experiments indicate that the quadratic simulations can be safely run with  $CFL = 0.5$ .

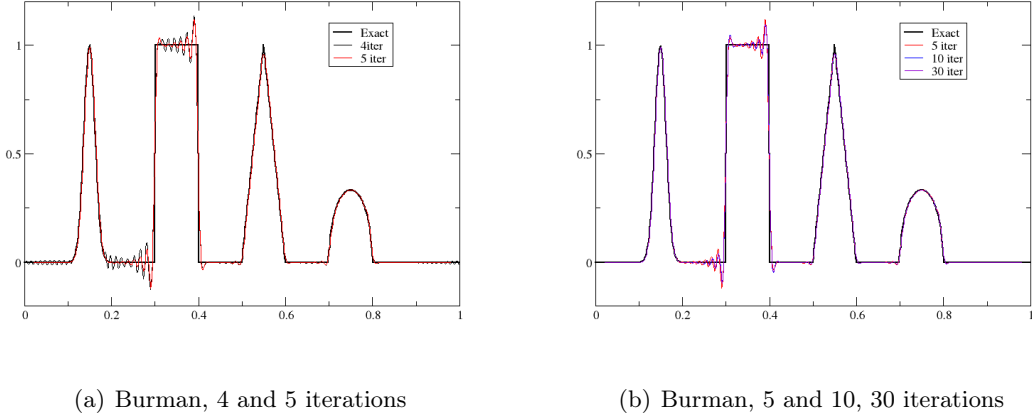


Figure 1: Results for the convection problem with initial conditions(41), at  $T = 8$ . The mesh has 200 cells (300 degrees of freedom). The results are obtained with the Burman residual.

Figure 4 displays the results for the  $\mathbb{P}^1$  approximation, while Figure 5 shows those obtained for the quadratic approximation. The baseline schemes are the SUPG and the Galerkin scheme with jumps.

In Figure 4, the same isolines are represented for the three results. We can see that after 10 rotations, the results of the Galerkin+jump scheme look pretty good despite the coarse resolution. The minimum and maximum are  $-0.012$  and  $0.762$ . For the SUPG results, after 1 rotation, the minimum/maximum are  $-0.004$  and  $1.02$ . After 2 rotations we have  $-0.047$  and  $1.02$ . This is better than what is obtained for Figure 4-(c), but the dispersive effects are much more important for the SUPG scheme as it can be seen on Figure 4-(b): this is why we have not shown further results for the SUPG/P1 case.

In Figure 5, we show similar results obtained with the quadratic approximation. Again, the Galerkin+jump method is way less dispersive than the SUPG (stopped after only one rotation this time). We have found that if we perform 4, 6 or 8 iterations of the defect correction, the quality of the SUPG improves a lot, but the cost becomes prohibitive with respect to the Galerkin+Jump method for which, after 10 rotations, the min/max are  $-0.0044$  and  $0.95$ . We also see that the solution improves a lot with respect to linear elements, for example in terms of min/max values. There is however some dispersion, if we compare with the exact solution. The table 2 display the error in the  $L^1$ ,  $L^2$  and  $L^\infty$  norm.

### 5.3 2D, non linear case: the KPP problem.

The second example is the KPP (Kurganov-Petrov-Petrova) test case, see [16]. The problem is described by

$$\begin{aligned} \frac{\partial u}{\partial t} + \operatorname{div} \mathbf{f}(u) &= 0 \\ u(\mathbf{x}, 0) &= u_0(\mathbf{x}) \end{aligned}$$

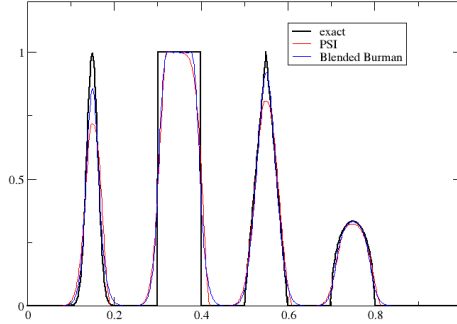


Figure 2: Results for the convection problem with initial conditions(41), at  $T = 8$  with quadratic elements. The results are obtained with PSI and Blended Burman schemes. Only 3 iterations are made

$\mathbf{f} = (\cos u, \sin u)$  and

$$u_0(\mathbf{x}) = \begin{cases} \frac{7}{2}\pi & \text{if } \|\mathbf{x}\| < 1 \\ \frac{\pi}{4} & \text{else.} \end{cases}$$

The flux is non convex, and we have two main difficulties:

- existence of composite waves, i.e. shock attached to fans, because the problem is not convex
- There exists a sonic points on  $\|x\| = 1$  (at  $\approx 112.5^\circ$ ).

Because of these two difficulties, if the scheme does not dissipate enough, then a shock wave is attached to the sonic point, this is not correct. If on contrary the scheme dissipates too much, then the solution can be blurred. This problem is more difficult than a standard problem with the flux  $\mathbf{f} = (\frac{u^2}{2}, u)$  which is convex.

Since the pure SUPG method is kind of disappointing, we have considered a spatial approximation using jump filtering, i.e. the approximation (40) with choices 1 and 2 for  $\beta$ . Some results are displayed on Figures 6 and 7. The mesh has been constructed by a mesh generator (gmesh). The second and third order simulations are done with exactly the same number of degrees of freedom (here 34353): the quadratic elements are subdivided into 4 linear elements. The second order solution uses a second order time discretisation, while the third order one a third order discretisation.

We notice that the second order solution presents crisper discontinuities. Indeed, in both case, the width of a discontinuity is of the order of 1.5 elements, but the size of the quadratic elements is twice as large as the one of the linear elements. Both solutions are correct, if one compares to published results, for example [16]. They have been obtained with choice 2 (Blended PSI). If choice 1 (PSI) is applied, the solution is not correct since the initial discontinuity stays attached, and then the composite wave cannot be created. In fact the Blended PSI is slightly more dissipative than the PSI one. We have not represented this solution to save space.



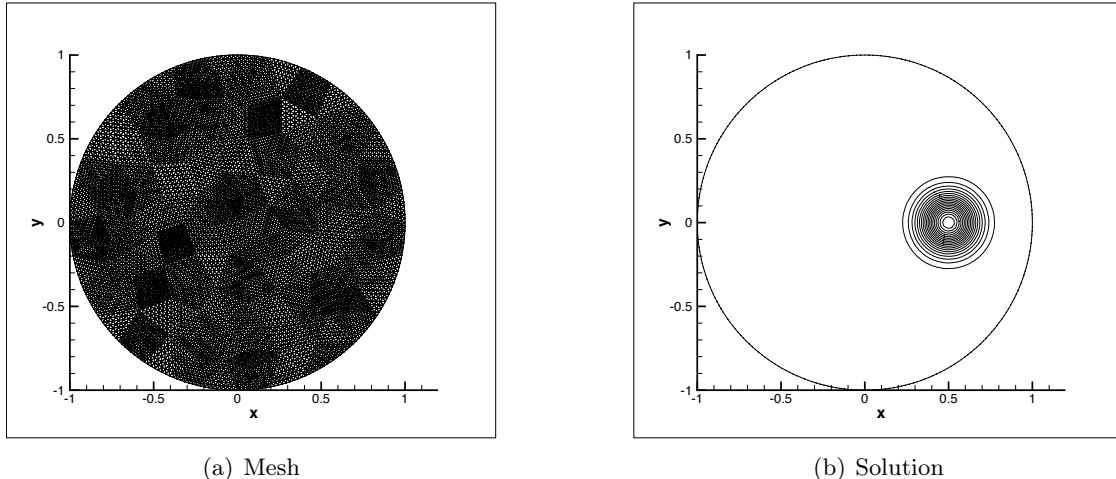


Figure 3: Exact solution after  $n$  rotations ( $n \in \mathbb{N}$ ) and plot of the degrees of freedoms.

## 6 Conclusions and perspectives

In this paper, we have shown how one can get rid-off mass matrix for continuous finite element method. This method relies on a iterative interpretation of the time stepping, in the spirit of Differed Correction method with the use of finite elements where each basis function has a stricktly positive mass. Some analysis indicates that one can get the expected accuracy, this is confirmed by the numerical results obtained on typical liner problems in one and two dimension. A formal extension is provided for non linear problems.

A natural perspective of this work is to apply the same approach to systems of conservation laws such a the Euler equations. Semi-implicit versions of the schemes can also be constructed in the same spirit.

### Acknowledgements.

The financial support of the SNF (under grant # 200021\_153604) is acknowledged. Many discussions with M. Ricchiuto (INRIA, Bordeaux Sud-Ouest, France) are acknowledged in the early stage of this work. S. Tokareva and P. Baccigalupi, both from the university of Zürich, are also acknowledged for their contributions in the early draft of this work. The contributions of A. Burbeau (CEA DEN, Saclay, France) are also acknowledged.

### References

- [1] R. Abgrall. Residual distribution schemes: current status and future trends. *Computer and Fluids*, 2006.
- [2] R. Abgrall. On a class of high order schemes for hyperbolic problems. In *Proceedings of the international Conference of Mathematicians*, volume II, Seoul, 2014.

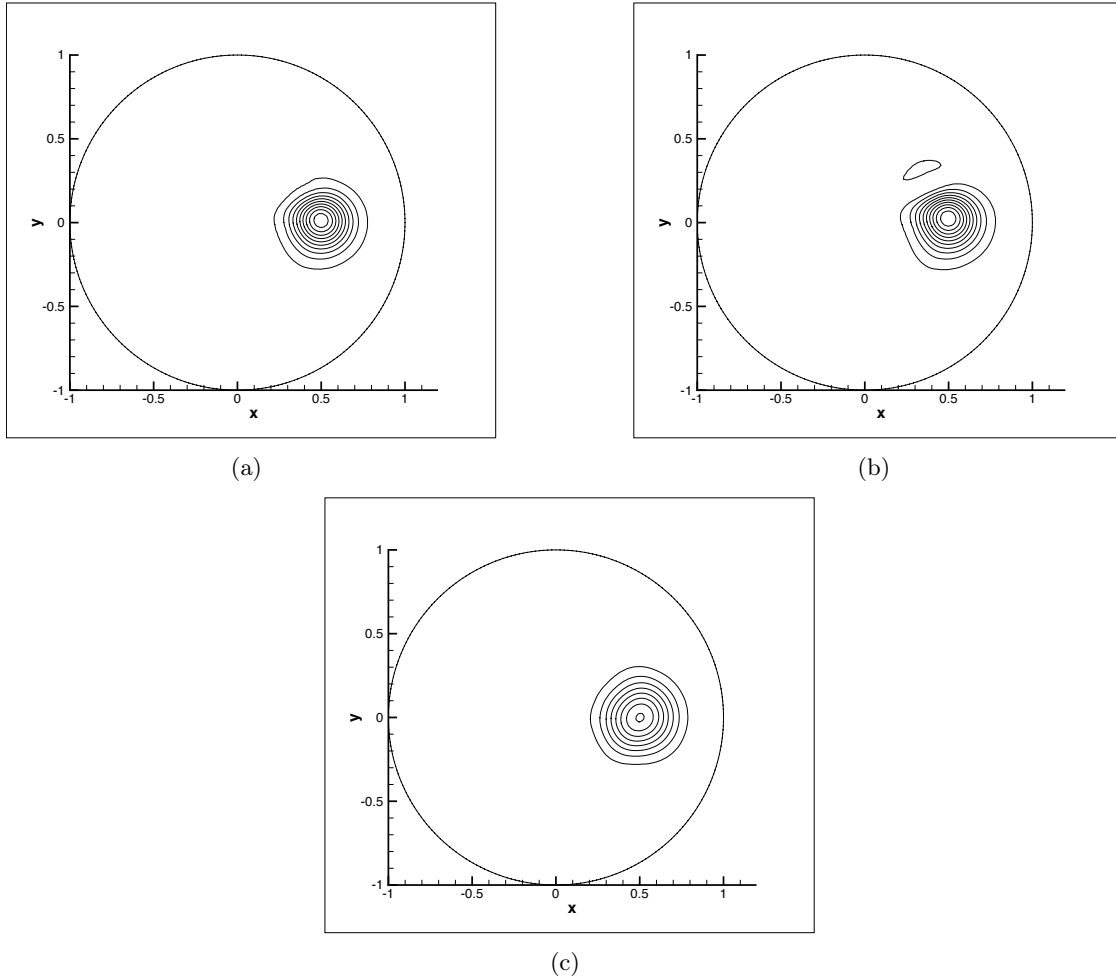


Figure 4: Results for the  $\mathbb{P}^1$  approximation: (a) with SUPG, after 1 rotation, (b) with SUPG after 2 rotations, (c) with Galerkin+Jump after 10 rotations. The same isolines are represented.

- [3] R. Abgrall, P. Bacigaluppi, and S. Tokareva. How to avoid mass matrix for linear hyperbolic problems. In B. Karasözen, M. Manguoglu, M. Tezer-Sezgin, S. Goktepe, and O. Ugur, editors, *Numerical Mathematics and Advanced Applications ENUMATH 2015*, volume Lecture Notes in Computational Sciences and Engineering, v112. Springer Verlag, 2016.
- [4] R. Abgrall and D. de Santis. Linear and non-linear high order accurate residual distribution schemes for the discretization of the steady compressible navier-stokes equations. *Journal of Computational Physics*, 283:329–359, 2015.
- [5] R. Abgrall, A. Larat, and M. Ricchiuto. Construction of very high order residual distribution schemes for steady inviscid flow problems on hybrid unstructured meshes. *J. Comput. Phys.*, 2011. <http://hal.inria.fr/inria-00464799/en>.

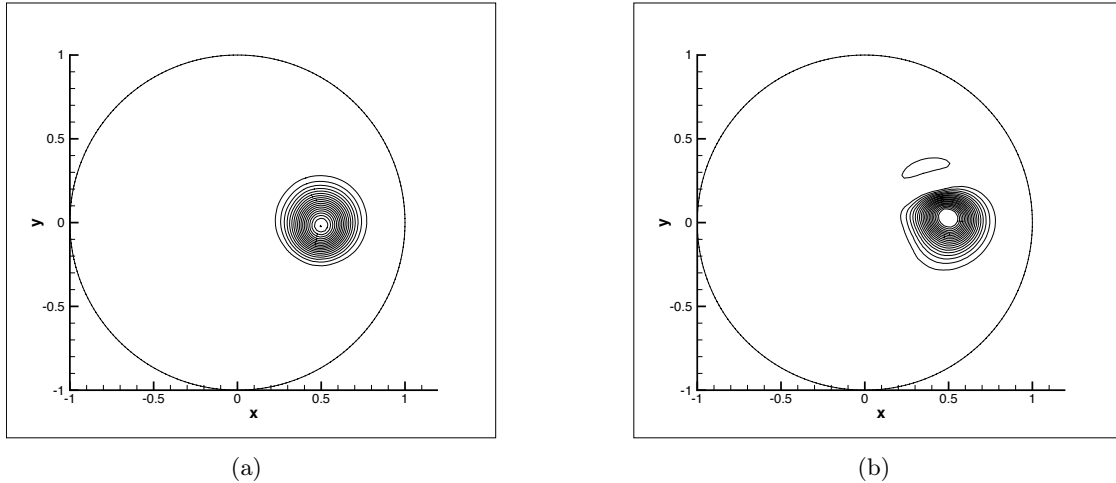


Figure 5: Results for the  $\mathbb{B}^2$  approximation: (a) with Galerkin+Jump after 10 rotations, (b) with SUPG after 1 rotations, (c). The same isolines are represented

- [6] Rémi Abgrall and Jirka Treflick. An example of high order residual distribution scheme using non-lagrange elements. *Journal of Scientific Computing*, 45(1-3):64–89, October 2010.
- [7] A.Ern and D. di Pietro. *Mathematical aspects of Discontinuous galerkin Methods*. Mathématiques et Applications. Springer, 2010.
- [8] A. Bourlioux, A.T. lyton, and M.L. Minion. High-order multi-implicit spectral defered correction methods for problem of reacting flow. *J. Comput. Phys.*, 2001.
- [9] E. Burman and P. Hansbo. Edge stabilization for Galerkin approximation of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg*, 193:1437–1453, 2004.
- [10] Manuel Quezada de Luna, Dmitri Kuzmin, Vladimir Z Tomov, Tzanio Kolev, Veselin A Dobrev, Robert N Rieben, and Robert Anderson. High-order local maximum principle preserving (mpp) discontinuous galerkin finite element method for the transport equation. *Journal of Computational Physics*, 2017. in press.
- [11] A. Dutt, L. Greengard, and V. Rokhlin. Spectral deferred correction methods for ordinary differential equations. *BIT Numerical Mathematics*, 2000.
- [12] E Godlewski and PA Raviart. *Hyperbolic systems of conservation laws*. Ellipses, February 1991.
- [13] Thomas J.R. Hughes and Michel Mallet. A new finite element formulation for computational fluid dynamics. III: The generalized streamline operator for multidimensional advective- diffusive systems. *Comput. Methods Appl. Mech. Eng.*, 58:305–328, 1986.
- [14] G.S.. Jiang and C.W. Shu. Efficient implementation of weighted ENO schemes. *Journal of Computational Physics*, 126:202–228, 1996.
- [15] L. Kridonova. Limites for high-order discontinuous Galerkin schemes. *Journal of Computational Physics*, 226:879–896, 2007.

$N_{dofs}$	$\mathcal{L}^1$	slope	$\mathcal{L}^2$	slope	$L^\infty$	slope
1236	$1.351 \cdot 10^{-1}$	—	$1.335 \cdot 10^{-1}$	—	$5.217 \cdot 10^{-1}$	—
4821	$2.997 \cdot 10^{-2}$	2.21	$4.207 \cdot 10^{-2}$	1.69	$1.967 \cdot 10^{-1}$	1.43
19041	$3.976 \cdot 10^{-3}$	2.94	$7.133 \cdot 10^{-3}$	2.58	$4.149 \cdot 10^{-2}$	2.26
75681	$6.710 \cdot 10^{-4}$	2.57	$1.217 \cdot 10^{-3}$	2.56	$7.063 \cdot 10^{-3}$	2.56

Second order

$N_{dofs}$	$\mathcal{L}^1$	slope	$\mathcal{L}^2$	slope	$L^\infty$	slope
4825	$2.508 \cdot 10^{-2}$	—	$3.056 \cdot 10^{-2}$	—	$1.161 \cdot 10^{-1}$	—
19041	$1.354 \cdot 10^{-3}$	4.24	$2.592 \cdot 10^{-3}$	3.54	$1.347 \cdot 10^{-2}$	3.13
75297	$1.094 \cdot 10^{-4}$	3.24	$2.003 \cdot 10^{-4}$	3.72	$1.137 \cdot 10^{-3}$	3.59
300993	$1.547 \cdot 10^{-5}$	2.82	$2.653 \cdot 10^{-5}$	2.91	$1.742 \cdot 10^{-4}$	2.70

Third order

Table 2: Rotation test case with quadratic+jump scheme. The error after one rotation is displayed for the initial condition  $u_0(\mathbf{x}) = e^{-\frac{\|\mathbf{x}-\mathbf{x}_0\|^2}{40}}$ . Here,  $h \approx \sqrt{N_{dofs}}$ .

- [16] Alexander Kurganov, Guergana Petrova, and Bojan Popov. Adaptive semidiscrete central-upwind schemes for nonconvex hyperbolic conservation laws. *SIAM J. Sci. Comput.*, 29(6):2381–2401, 2007.
- [17] Y. Liu, C.-W. Shu, and M. Zhang. Strong stability preserving property of the deferred correction time discretisation. *Journal of Computational Mathematics*, 26(5):633–656, 2008.
- [18] M.L. Minion. Semi-implicit spectral deferred correction methods for ordinary differential equations. *Communication in Mathematical Physics*, 1(3):471–500, 2003.
- [19] Mario Ricchiuto and Rémi Abgrall. Explicit runge-kutta residual-distribution schemes for time dependent problems. *Journal of Computational Physics*, 229(16):5653–5691, 2010.
- [20] R. D. Richtmyer and K. W. Morton. *Difference Methods for Initial-Value Problems*. Interscience, New-York, 1967.

## A Technical results

**Lemma A.1.** Assume that  $K$  is convex and its aspect ratio is bounded by a constant  $C$ . If  $v \in \mathbb{P}_r(K)$ , and  $v = \sum_{\sigma \in K} v_\sigma \varphi_\sigma$ , then

$$\sum_{\sigma} |v_\sigma - v_{\sigma'}| \leq C_K \sum_{\sigma} |v(\sigma) - v(\sigma')|$$

where  $C_K$  is the  $L^\infty$  norm of the inverse of the matrix  $(\varphi_\sigma(\sigma'))_{\sigma, \sigma'}$ . and

$$h_K \sum_K |v_\sigma| \leq C_K \|v\|_{2,K}$$

where  $C_K$  only depends on  $K$  via  $C$ .

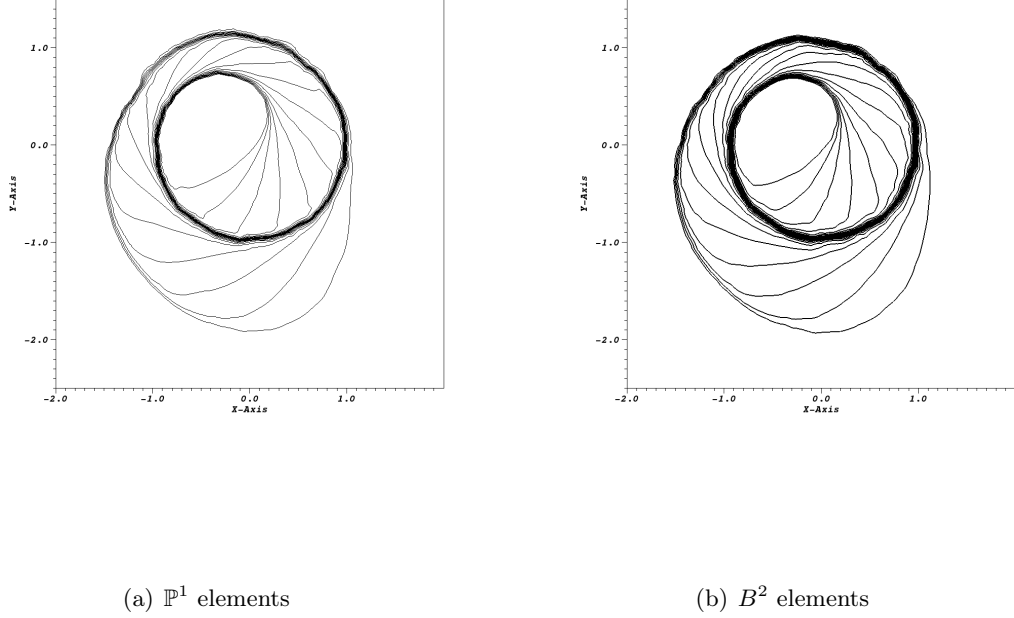


Figure 6: KPP problem, solution at  $t = 1$  for the second and third order scheme.

*Proof.* We have  $v(\sigma) = \sum_{\sigma' \in K} v(\sigma') \varphi_{\sigma'}(\sigma)$ , so that

$$\sum_{\sigma \in K} |v(\sigma)| \leq C_1 \sum_{\sigma \in K} |v(\sigma)|^2 \leq C_1 \|A^{-1}\|_2 \int_K v^2(\mathbf{x}) d\mathbf{x}$$

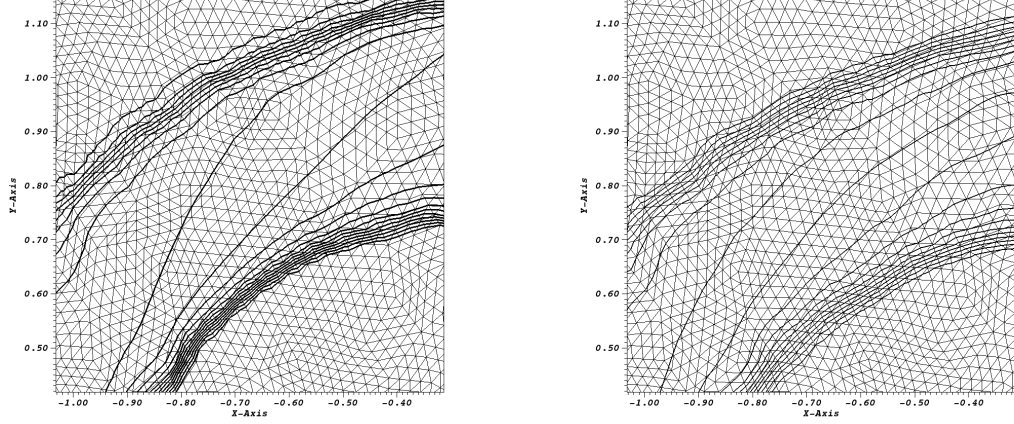
where  $A$  is the matrix  $A = (\int_K \varphi_{\sigma'} \varphi_{\sigma})_{\sigma, \sigma' \in K}$ , and  $C_1$  is the square root of the number of degrees of freedom in  $K$ .

By a scaling argument,  $\|A^{-1}\|_2 \leq C_K h_K^{-1}$  where  $C_K$  depends on the aspect ratio of  $K$ . Hence,

$$\sum_{\sigma \in K} |v_{\sigma} - v_{\sigma'}| \leq C_K \sum_{\sigma \in K} |v(\sigma) - v(\sigma')| \leq \frac{C_K}{h_K} \int_K |v(\mathbf{x}) - v(\sigma')| d\mathbf{x}$$

where  $C_k$  is the  $L^\infty$  norm of the matrix  $(\varphi_{\sigma'})_{\sigma, \sigma' \in K}$ . We have:

$$v(\mathbf{x}) - v(\sigma') = \int_0^1 \nabla v((1-s)\mathbf{x} + s\sigma') \cdot (\mathbf{x} - \sigma) ds,$$



(a)  $\mathbb{P}^1$  elements

(b)  $B^2$  elements

Figure 7: KPP problem, solution at  $t = 1$  for the second and third order scheme. Zoom of the solution, the degrees of freedom are represented.

so that

$$\begin{aligned}
\int_K |v(\mathbf{x}) - v(\sigma')| d\mathbf{x} &\leq \int_K \int_0^1 \|\nabla v((1-s)\mathbf{x} + s\sigma')\| \|\mathbf{x} - \sigma'\| ds d\mathbf{x} \\
&\leq h_K \int_K \left( \int_0^1 \|\nabla v((1-s)\mathbf{x} + s\sigma')\|^2 ds \right)^{1/2} d\mathbf{x} \\
&\leq h_K \left( \int_K \left( \int_0^1 \|\nabla v((1-s)\mathbf{x} + s\sigma')\|^2 ds \right) d\mathbf{x} \right)^{1/2}
\end{aligned}$$

since  $s \mapsto \sqrt{s}$  is concave. Using Fubini, we then have

$$\int_K \left( \int_0^1 \|\nabla v((1-s)\mathbf{x} + s\sigma')\|^2 ds \right) d\mathbf{x} = \int_{K \times [0,1]} \|\nabla v((1-s)\mathbf{x} + s\sigma')\|^2 = \int_K \|\nabla v(\mathbf{x})\|^2 d\mathbf{x}$$

because  $K$  is convex.

Collecting all the pieces, we get:

$$\sum_{\sigma \in K} |v_\sigma - v_{\sigma'}| \leq C_K \|\nabla v\|_{\mathcal{L}^2(K)}$$

where  $C_K$  only depends on the aspect ratio of  $K$ .

□

## B Some properties of non linear RDS schemes.

This annex is devoted to the justification of some fact stated in section 4.4, namely that the  $\mathcal{L}^2$  operator, i.e. for each element and each sub-time step  $p$ ,

$$\int_K \psi_\sigma(V_p^{(l)} - V_0)dx + \int_K \psi_\sigma \left( \sum_{k=1}^{r+1} \theta_{k,r+1} \mathbf{f}(V_k^{(l)}) \right)$$

can write

$$|K| \sum_{\sigma' \in K} a_{\sigma\sigma'}^K (V_{p,\sigma'}^{(l)} - u_{\sigma'}) + \Delta t \sum_{k=1}^{r+1} \theta_{k,r+1} \left\{ \sum_{\sigma' \in K} c_{\sigma\sigma'}^k (V_{k,\sigma}^{(l)} - V_{k,\sigma'}^{(l)}) \right\}.$$

with

- $a_{\sigma\sigma}^K = \gamma_\sigma^K \frac{|K|}{\#DOF}$  with  $\gamma_\sigma^K \in [0, 1]$  and  $\#DOF$  is the number of degrees of freedom in  $K$
- and  $c_{\sigma\sigma'}^k \geq 0$ .

We consider the following kind of nonlinear RDS. The Galerkin residuals are defined by

$$\Phi_\sigma^{K,Gal} = \int_{\partial K} \varphi_\sigma \left( \sum_{k=1}^{r+1} \theta_{k,r+1} \mathbf{f}(V_k^{(l)}) \right) \cdot \mathbf{n} - \int_K \nabla \varphi_\sigma \left( \sum_{k=1}^{r+1} \theta_{k,r+1} \mathbf{f}(V_k^{(l)}) \right), \quad (42)$$

from this one writes a ‘‘Rusanov’’ residual:

$$\Phi_\sigma^{K,Rus} = \frac{|K|}{\#K} \left( (V_{p,\sigma}^{(l)} - V_{0,\sigma}) + \Delta t \Phi_\sigma^{K,Gal} + \Delta t \alpha_K \left( \left\{ \sum_{k=1}^{r+1} \theta_{k,r+1} V_k^{(l)} \right\} - \bar{V} \right) \right)$$

with

$$\bar{V} = \frac{1}{\#K} \left( \sum_{\sigma \in K} \sum_{k=1}^{r+1} \theta_{k,r+1} V_k^{(l)} \right)$$

and  $\alpha_K$  larger than the maximum of the spectral radii of the Jacobians of the flux evaluated at the states  $V_k^{(l)}$ , or even larger. Then one forms

$$\Phi_\sigma^{K,*} = \beta_\sigma^K \Phi_{xt}^K \quad (43)$$

where the total residual  $\Phi^K$  is defined by

$$\Phi_{xt}^K = \int_K (V_p^{(l)} - V_0)dx + \int_{\partial K} \left( \sum_{k=1}^{r+1} \theta_{k,r+1} \mathbf{f}(V_k^{(l)}) \right) \cdot \mathbf{n} (= \sum_{\sigma \in K} \Phi_\sigma^{K,Gal})$$

and  $\beta_\sigma^K$  by

$$\beta_\sigma^K = \frac{\max(0, \frac{\Phi_\sigma^{K,Rus}}{\Phi_{xt}^K})}{\sum_{\sigma' \in K} \max(0, \frac{\Phi_{\sigma'}^{K,Rus}}{\Phi_{xt}^K})}. \quad (44)$$

Let us prove now that

$$a_{\sigma\sigma}^K = \gamma_\sigma^K \frac{|K|}{\#DOF} \text{ with } \gamma_\sigma^K \in [0, 1] \text{ and } \#DOF \text{ is the number of degrees of freedom in } K. \quad (45)$$

and

$$c_{\sigma\sigma'}^k \geq 0. \quad (46)$$

*Proof.* We note that

$$\beta_\sigma^K \Phi_{xt}^K = \gamma_\sigma^K \Phi_\sigma^{K,Rus}$$

with

$$\gamma_\sigma^K = \begin{cases} 0 & \text{if } \max(0, \frac{\Phi_\sigma^{K,Rus}}{\Phi_{xt}^K}) = 0 \\ \frac{1}{\sum_{\sigma' \in K} \max(0, \frac{\Phi_{\sigma'}^{K,Rus}}{\Phi_{xt}^K})} & \text{else} \end{cases}$$

Since  $\sum_{\sigma \in K} \Phi_\sigma^{K,Rus} = \Phi_{xt}^K$ , we have that:

$$\sum_{\sigma' \in K} \max(0, \frac{\Phi_{\sigma'}^{K,Rus}}{\Phi_{xt}^K}) + \sum_{\sigma' \in K} \in (0, \frac{\Phi_{\sigma'}^{K,Rus}}{\Phi_{xt}^K}) = \sum_{\sigma' \in K} \frac{\Phi_{\sigma'}^{K,Rus}}{\Phi_{xt}^K} = 1,$$

so that

$$\sum_{\sigma' \in K} \max(0, \frac{\Phi_{\sigma'}^{K,Rus}}{\Phi_{xt}^K}) \geq 1$$

and then  $\gamma_\sigma^K \in [0, 1]$ . We get the first property (45).

The second one (46) comes from the very definition of  $\beta_\sigma^K$ . □

**Remark B.1.** *In many practical applications, the residual that is considered is not (43) but*

$$\Phi_\sigma^{K,*} = \beta_\sigma^K \Phi_{xt}^K + \sum_{\text{edges of } K} h_K^2 \Gamma \int_e [\nabla \varphi_\sigma] \left[ \nabla \left( \sum_{k=1}^{r+1} \theta_{k,r+1} V_k^{(l)} \right) \right] \quad (47)$$

with  $\beta_\sigma^K$  defined as (44). Then (45) is still true because the term

$$\int_{\partial K} \left( \sum_{k=1}^{r+1} \theta_{k,r+1} \mathbf{f}(V_k^{(l)}) \right) \cdot \mathbf{n} (= \sum_{\sigma \in K} \Phi_\sigma^{K,Gal})$$

does not contain any time increment.

In some other, we modify the definition of the Rusanov residual into

$$\Phi_\sigma^{K,Rus} = \Phi_\sigma^{K,Gal} + \Delta t \alpha_K \left( \left\{ \sum_{k=1}^{r+1} \theta_{k,r+1} V_k^{(l)} \right\} - \bar{V} \right)$$

with now

$$\Phi_\sigma^{K,Gal} = \int_K (V_p^{(l)} - V_0) \varphi_\sigma dx \Delta t \int_{\partial K} \varphi_\sigma \left( \sum_{k=1}^{r+1} \theta_{k,r+1} \mathbf{f}(V_k^{(l)}) \right) \cdot \mathbf{n} - \Delta t \int_K \nabla \varphi_\sigma \left( \left( \sum_{k=1}^{r+1} \theta_{k,r+1} \mathbf{f}(V_k^{(l)}) \right) \right)$$

and we consider

$$\Phi_\sigma^{K,*} = (1 - \ell) \left\{ \Phi_\sigma^{K,Gal} + \sum_{\text{edges of } K} h_K^2 \Gamma \int_e [\nabla \varphi_\sigma] \left[ \nabla \left( \sum_{k=1}^{r+1} \theta_{k,r+1} V_k^{(l)} \right) \right] \right\} + \ell \Phi_\sigma^{K,Rus} \quad (48)$$

with

$$\ell = \frac{|\Phi^K|}{\sum_{\sigma \in K} |\Phi_\sigma^{K,Rus}|}. \quad (49)$$

Then none of the properties hold formally true but we get a maximum principle experimentally.