High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices

R Abgrall

High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices.

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

July 8, 2017

Abstract

When integrating unsteady problems using globally continuous representation of the solution, as for 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 other 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 [4] and [26].

1 Introduction

We are interested in the numerical approximation of the hyperbolic problem

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

with the initial condition and boundary conditions, by using a globally continuous approximation of $\mathbf{u}$ as in a finite element methods[1]

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

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

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

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 div $\mathbf{f}$ and M is a mass matrix. In the case of continuous elements, this matrix is sparse but is not block diagonal,

[1] In this paper we also show that some of the finite element methods for approximating (1) can beneficiate of the techniques developped here.

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 [2], 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 question of avoiding the inversion of mass matrix has of course motivated many researchers over the years. One may mention the work of Löhner et al. [24], Donea et al. [13], Guermond et al.[17], Cohen et al. [11], Jund et al. [20], or Ricchiuto et al. [26]. We believe that this contribution is different in spirit with [24, 13, 17] in that our method is not really an iterative method since the number of steps is controlled, but it can be seen as a generalisation and a more systematic version of [26]. In [11, 20], high-order trial spaces are defined and they are compatible with a genuine mass-lumping approximation. However, the CFL condition become more and more strict as the degree increases, while in this paper the time step behaves like $1/k$ where k is the polynomial degree of the trial space. For the sake of completeness and show some connections and difference to more standard methods, we also show, in an appendix, that some Runge-Kutta ODE integrators can be put in a deferred correction framework, and how to use, in principle, multi-step methods.

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 final justification of the temporal scheme with respect to more classical strategies is presented in section 6. 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, recall some basic facts about Residual distribution schemes and provide a reinterpretation of some Runge-Kutta schemes as DeC schemes.

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

2 Continuous problem setting

We consider

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

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

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

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 $\overline{\nabla_u \mathbf{f}_\mathbf{n}}(a,b)$ such that

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

This is a reminscence 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

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

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

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

We introduce the space $C^1_{0,t}(\Omega \times \mathbb{R}^+)$ of functions with compact support in time,

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

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

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

This is a direct generalization of what is done for $\mathbf{f}(\mathbf{u}) = \beta\mathbf{u}$ when β is constant. See [8] for more. For existence and uniqueness results, see for example [16].

3 Short review of DeC schemes for ODEs

We consider the problem:

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

with suitable conditions on f to guarantee 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 Defect Correction scheme (DeC). Let us write formally the (discrete) problem to solve as $\mathcal{L}^2 = 0$. Recall we assume it to be implicit, not necessarily because the problem is stiff, but because the formulation of the numerical problem is of this type. Hence it is a complicated one. In addition, f in (4) is non linear and its derivative may be difficult to estimate accurately. 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 a way $\mathcal{L}^1$ as a kind of preconditionner. 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 [14, 9, 25], see also the references therein.

Let us now be more specific. 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:

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

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

Let us be given a time step Δt 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: $t_{n,\ell} = t_n + \xi_\ell \Delta t$, $\ell = 0, \ldots, r$ and $0 = \xi_0 < \ldots < \xi_\ell < \xi_{\ell+1} < \ldots < \xi_r = 1$. 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_ℓ, $\ell = 0, \ldots, r$

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

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

  This is the forward Euler method.

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

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

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

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

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

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

we can rewrite (8) as

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

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

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

  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 numbers $\theta_{l,m}$ 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:

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

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 $y^{n+1} = y_r^{(M)}$.

In practice, provided conditions recalled in section 4.3.1 are met, at most M iterations are needed to have a $\min(M+1, r+1)$th order accurate approximation of the solution. In addition, the accuracy of the solution increases by one order for each iteration up to order $M+1$.

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

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

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

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

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

4 Application to the convection problem

We are interested in solving

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

subjected to $u(x, 0) = u_0(x)$ for $x \in \Omega \subset \mathbb{R}^d$ and the boundary conditions (2c). We assume $d = 2$ to simplify the text, but the generalisation to $d = 3$ (or $d = 1$) is straightforward. We are given a triangulation of Ω so that Ω is covered by the elements of the triangulation which is assumed to be conformal. They are genericaly denoted by K and are assumed to be simplices. The boundary $\partial\Omega$ is covered by segments (faces in 3D) that we genericaly denote by Γ. The internal faces are genericaly denoted by e and their collection is the squeleton of the triangulation.

In each element, we assume that the solution is approximated by a polynomial of degree r and that the approximation is globally continuous in Ω. Let us denote this by u^h. The function u^h is fully defined by its control parameter u_σ at all the degrees of freedom σ. 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_σ as the value of u^h at σ and thus as the φ_σ 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 [23].

4.1 Spatial approximation, link to continuous finite elements

In order to integrate the steady version of (13) on a domain $\Omega \subset \mathbb{R}^d$ with the boundary conditions (2c), on each element K and any degree of freedom $\sigma \in \mathcal{S}$ belonging to K, we define residuals

$\Phi_\sigma(u^h)^K$. Following [6, 5, 3], they are assumed to satisfy the following conservation relations: For any element K,

$$\sum_{\sigma\in K}\Phi_\sigma^K(u_h)=\int_{\partial K}\mathbf{f}(u^h)\cdot\mathbf{n}. \tag{14}$$

Similarly, we consider residuals on the boundary elements Γ. On any such Γ, for any degree of freedom $\sigma\in\mathcal{S}\cap\Gamma$, we consider boundary residuals $\Phi_\sigma^\Gamma(u^h)$ that will satisfy the conservation relation

$$\sum_{\sigma\in\Gamma}\Phi_\sigma^\Gamma(u_h)=\int_\Gamma\big(\mathcal{F}_\mathbf{n}(u^h,g)-\mathbf{f}(u^h)\cdot\mathbf{n}\big). \tag{15}$$

One this is done, the discretisation of (13) is achieved via: for any $\sigma\in\mathcal{S}$,

$$\sum_{K\subset\Omega,\sigma\in K}\Phi_\sigma^K(u^h)+\sum_{\Gamma\subset\partial\Omega,\sigma\in\Gamma}\Phi_\sigma^\Gamma(u^h)=0. \tag{16}$$

In (16), the first term represents the contribution of the internal elements. The second exists if $\sigma\in\Gamma$ and represents the contribution of the boundary conditions.

Indeed, the formulation (16) is very natural. Consider a variational formulation of the steady version of (13):

$$\text{find } u^h\in V_h \text{ such that for any } v^h\in V^h, a(u^h,v^h)=0.$$

We give two example to fix the ideas:

- the SUPG [18] variational formulation:

$$\begin{aligned}a(u^h,v^h):=&-\int_\Omega\nabla v^h\cdot\mathbf{f}(u^h)\,d\mathbf{x}+\sum_{K\subset\Omega}h_k\int_K\nabla\mathbf{f}(u^h)\nabla v^h\tau_K\nabla\mathbf{f}(u^h)\cdot\nabla u^h d\mathbf{x}\\&+\int_{\partial\Omega}v^h\big(\mathcal{F}_\mathbf{n}(u^h,g)-\mathbf{f}(u^h)\cdot\mathbf{n}\big)d\partial\Omega.\end{aligned} \tag{17}$$

- The Galerkin scheme with jump stabilization, see [10] for details: If e represents a generic *internal* edge, defining for any function ψ the jump $[\nabla\psi]=\nabla\psi_{|K}-\nabla\psi_{|K^+}$, we have

$$\begin{aligned}a(u^h,v^h):=&-\int_\Omega\nabla v^h\cdot\mathbf{f}(u^h)\,d\mathbf{x}+\sum_{e\subset\Omega}h_k^2\int_e\big[\nabla v^h\big]\cdot\big[\nabla u^h\big]\,de\\&+\int_{\partial\Omega}v^h\big(\mathcal{F}_\mathbf{n}(u^h,g)-\mathbf{f}(u^h)\cdot\mathbf{n}\big)d\partial\Omega.\end{aligned} \tag{18}$$

Using the fact that the basis functions that span V_h have a *compact* support, then both schemes can be rewritten in the form(16) with the following expression for the residuals:

- For the SUPG scheme (17), the residual are defined by

$$\Phi_\sigma^\mathbf{x}(u^h)=\int_{\partial K}\varphi_\sigma\mathbf{f}(u^h)\cdot\mathbf{n}-\int_K\nabla\varphi_\sigma\cdot\mathbf{f}(u^h)+h_K\int_K\Big(\nabla_u\mathbf{f}(u^h)\cdot\nabla\varphi_\sigma\Big)\tau\Big(\nabla_u\mathbf{f}(u^h)\cdot\nabla u^h\Big) \tag{19}$$

  with $\tau>0$.

- For the Galerkin scheme with jump stabilization (18), the residuals are defined by:

$$\Phi_\sigma^{\mathbf{x}}(u^h) = \int_{\partial K} \varphi_\sigma \mathbf{f}(u^h)\cdot\mathbf{n} - \int_K \nabla\varphi_\sigma\cdot\mathbf{f}(u^h) + \sum_{\text{edges of } K} \Gamma h_e^2 \int_e [\nabla u]\cdot[\nabla\varphi_\sigma] \tag{20}$$

  with $\Gamma > 0$. Here, since the mesh is conformal, any internal edge e (or face in 3D) is the intersection of the element K and an other element denoted by K^+.

- The boundary residuals are

$$\Phi_\sigma^{\Gamma}(u^h) = \int_\Gamma \big(\mathcal{F}_{\mathbf{n}}(u^h, g) - bF(u^h)\cdot\mathbf{n}\big) \tag{21}$$

All these residuals satisfy the relevant conservation relations, namely (14) or (15), depending if we are dealing with element residuals or boundary residuals.

At this level, we are just rephrasing classical finite element schemes into a purely numerical framework. However, considering further the pure numerical point of view, we can go further and define schemes that have no clear variational formulation. These are the limited Residual distributive scheme (RDS), see [6, 5, 3], namely

$$\Phi_\sigma^K = \beta_\sigma \int_{\partial K} \mathbf{f}(u^h)\cdot\mathbf{n} + h_K \int_K \Big(\nabla_u \mathbf{f}(u^h)\cdot\nabla\varphi_\sigma\Big)\tau\Big(\nabla_u \mathbf{f}(u^h)\cdot\nabla u^h\Big) \tag{22}$$

or

$$\Phi_\sigma^K(u^h) = \beta_\sigma \int_{\partial K} \mathbf{f}(u^h)\cdot\mathbf{n} + \sum_{\text{edges of } K} \Gamma h_e^2 \int_e [\nabla u]\cdot[\nabla\varphi_\sigma]. \tag{23}$$

where the parameters β_σ are defined to guarantee conservation,

$$\sum_{\sigma\in K} \beta_\sigma = 1$$

and such that (22) without the streamline term and (23) without the jump terms satisfy a discrete maximum principle. The stream line term and jump term are introduced because one can easily see that spurious modes may exist, but their role is very different compared to (19) and (20) where they are introduced to stabilize the Galerkin schemes: if formaly the maximum principle is violated, experimentally the violation is extremely small, if not inexistant. See [1, 6] for more details. Possible values of β_σ are described in the appendix C.

Remark 4.1. *We notice that the SUPG and the limited RDS residuals write as*

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

where:

- $\Psi_\sigma = \varphi_\sigma + h_K\big(\nabla_u \mathbf{f}(u^h)\cdot\nabla\varphi_\sigma\big)\tau$ *for the SUPG scheme,*
- *for the limited RDS* (19)*, we take*

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

This implies that formally at least, the exact solution cancels the SUPG and RDS residuals. In the case of the stabilisation by jumps (23)*, we can only write that*

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

with

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

We note that $\sum_{\sigma\in K} R_\sigma = 0$.

The additional term R_σ *is not zero, except for the exact solution unless this solution has continuous normal gradients, see [10] for more details. In any case, we note that*

$$\sum_{\sigma\in K} \psi_\sigma = 1. \tag{24}$$

4.2 Formulation for unsteady problems

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

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

and the approximate by

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

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^σ is the control parameter at the degree of freedom σ: $V = \sum_{\sigma\in\mathcal{S}} V^\sigma\varphi_\sigma$. Then,

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

$$\mathcal{L}_\sigma^1(V_1,\ldots,V_r) = \begin{pmatrix} |C_\sigma|(V_r^\sigma - V_{r-1}^\sigma) + \sum_{K,\sigma\in K}\int_{t_{n,0}}^{t_{n,r}} \mathcal{I}_0\big(\Phi_\sigma^{\mathbf{x}}(V),s)\big)ds \\ \vdots \\ |C_\sigma|(V_1^\sigma - V_0^\sigma) + \sum_{K,\sigma\in K}\int_{t_{n,0}}^{t_{n,1}} \mathcal{I}_0\big(\Phi_\sigma^{\mathbf{x}}(V),s)\big)ds \end{pmatrix} \tag{25a}$$

  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

$$\mathcal{L}_\sigma^2(V_1,\ldots,V_r) = \begin{pmatrix} \sum_{K,\sigma\in K}\left(\int_K \Psi_\sigma\big(V_r - V_0\big);d\mathbf{x} + \int_{t_{n,0}}^{t_{n,r}} \mathcal{I}_{r+1}\big(\Phi_\sigma^{\mathbf{x}}(V),s\big)\ ds\right) \\ \vdots \\ \sum_{K,\sigma\in K}\left(\int_K \Psi_\sigma\big(V_1 - V_0\big)\ d\mathbf{x} + \int_{t_{n,0}}^{t_{n,1}} \mathcal{I}_{r+1}\big(\Phi_\sigma^{\mathbf{x}}(V),s\big)\ ds\right) \end{pmatrix} \tag{25b}$$

In (25b), Ψ_σ is any of the test function defined in the previous paragraph. In (25) and (25b), $\Phi_\sigma^{\mathbf{x},K}$ represents any of the residual defined in the previous paragraph. Here we have put a superscript $\mathbf{x}$ to indicate these are spatial residual.

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

$$\mathcal{L}^1_{\sigma,K}(V_1,\dots,V_r) = \begin{pmatrix} |C_{\sigma,K}|(V_r^\sigma - V_{r-1}^\sigma) + \displaystyle\int_{t_{n,0}}^{t_{n,r}} \mathcal{I}_0\big(\Phi_\sigma^{\mathbf{x}}(V),s)\big)\ ds \\ \vdots \\ |C_{\sigma,K}|(V_1^\sigma - V_0^\sigma) + \displaystyle\int_{t_{n,0}}^{t_{n,1}} \mathcal{I}_0\big(\Phi_\sigma^{\mathbf{x}}(V),s)\big)\ ds \end{pmatrix}, \tag{26a}$$

and

$$\mathcal{L}^2_{\sigma,K}(V_1,\dots,V_r) = \begin{pmatrix} \displaystyle\int_K \Psi_\sigma\big(V_r - V_0\big)\ d\mathbf{x} + \int_{t_{n,0}}^{t_{n,r}} \mathcal{I}_{r+1}\big(\Phi_\sigma^{\mathbf{x}}(V),s\big)\ ds \\ \vdots \\ \displaystyle\int_K \Psi_\sigma\big(V_1 - V_0\big)\ d\mathbf{x} + \int_{t_{n,0}}^{t_{n,1}} \mathcal{I}_{r+1}\big(\Phi_\sigma^{\mathbf{x}}(V),s\big)\ ds \end{pmatrix} \tag{26b}$$

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

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

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

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

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

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

$$(\mathcal{L}^1_{\sigma,K,t})_p = |C_{\sigma,K}|(V_{p+1} - V_p), \qquad (\mathcal{L}^2_{\sigma,K,t})_p = \int_K \Psi_\sigma\big(V_p - V_0\big)dx \tag{27}$$

and

$$(\mathcal{L}^1_{\sigma,K,\mathbf{x}})_p = \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_0\big(\Phi_\sigma^{\mathbf{x}}(V,s)\big)ds, \qquad (\mathcal{L}^2_{\sigma,K,\mathbf{x}})_p = \int_{t_{n,0}}^{t_{n,p}} \mathcal{I}_{r+1}\big(\Phi_\sigma^{\mathbf{x}}(V,s)\big)ds. \tag{28}$$

The operator $\mathcal{L}^2$ is space time consistent, while $\mathcal{L}^1$ is not, and thus can be only first order in time.

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:

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

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

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

Then, we evaluate u^{n+1}_σ as in the ODE case:

- Evaluate $V^{(0)} = (V_1^{(0)},\dots,V_{r+1}^{(0)})$ as the solution of $\mathcal{L}^1_\sigma(V^{(0)}) = 0$. 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

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

More explicitly, we have:

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

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

- Correction step $\#m$:

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

i.e.

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

Remark 4.2 (Other choice for $\mathcal{L}^1$)**.** *An other possibility is to consider $\mathcal{L}^1_\sigma$ defined by*

$$\mathcal{L}^1_\sigma(V_1,\dots,V_{r+1}) = \begin{pmatrix}|C_\sigma|(V_{r+1}-V_0) + \Delta t\ \Phi^{\mathbf{x}}_\sigma(V_0)\\ |C_\sigma|(V_r-V_0) + \xi_r\ \Delta t\ \Phi^{\mathbf{x}}_\sigma(V_0)\\ \vdots\\ |C_\sigma|(V_1-V_0) + \xi_1\ \Delta t\ \Phi^{\mathbf{x}}_\sigma(V_0)\end{pmatrix} \tag{29}$$

with $\xi_p = \alpha_1 + \ldots + \alpha_p$. This leads to a simpler implementation.

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 trial and test functions to get a meaningful approximation.

4.3 Analysis

4.3.1 Introduction

It is well known that

Proposition 4.3. *If two operators $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$ defined on $\mathbb{R}^m$, which depend of a parameter Δ, are such that:*

1. *We assume there exists a unique $U^\star_\Delta$ such that $\mathcal{L}^2_\Delta(U^\star_\Delta) = 0$.*
2. *There exists $\alpha_1 > 0$ independent of Δ such that for any U, V,*

$$\alpha_1 ||U - V|| \leq ||\mathcal{L}^1_\Delta(U) - \mathcal{L}^1_\Delta(V)||, \tag{30}$$

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

$$\left|\left|\big(\mathcal{L}^1_\Delta(U) - \mathcal{L}^2_\Delta(U)\big) - \big(\mathcal{L}^1_\Delta(U) - \mathcal{L}^2_\Delta(U)\big)\right|\right| \leq \alpha_2 \Delta ||U - V||. \tag{31}$$

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

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 Δ to simplify the text. Let us denote by $U^\star$ the solution of $\mathcal{L}^2(U^\star) = 0$ We obviously have $\mathcal{L}^1(U^\star) = \mathcal{L}^1(U^\star) - \mathcal{L}^2(U^\star)$, so that

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

so that

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

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 (30) ?
- Under which conditions do we have

$$||\mathcal{L}^1 - \mathcal{L}^2|| = O(\Delta t) + O(h). \tag{32}$$

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

4.3.2 Choice of norms

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

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

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

$$||\Phi|| = \sup_{u \in H^1(\Omega)} \frac{\langle u, \Phi \rangle}{||u||_{H^1(\Omega)}} \tag{33a}$$

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

$$||\Phi|| = \max_{i=1,\ldots,m} ||\Phi_m||. \tag{33b}$$

4.3.3 Galerkin scheme

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

$$\Phi(V_{l,K})^{G,\mathbf{x}}_\sigma = -\int_K \nabla \cdot \varphi_\sigma \mathbf{f}(V_{l,K}) d\mathbf{x} + \int_{\partial K} \varphi_\sigma \mathbf{f}(V_{l,K}) \cdot \mathbf{n}. \tag{34}$$

We see that

$$\sum_{\sigma \in K} \Phi(V_{l,K})^{G,\mathbf{x}}_\sigma = \int_K \mathbf{f}(V_{l,K}) \cdot \mathbf{n} = \sum_{\sigma \in K} \Phi(V_{l,K})^{\mathbf{x}}_\sigma \tag{35}$$

for any of the spatial residuals defined above.

4.3.4 Coercivity of $\mathcal{L}^1$

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

Lemma 4.4. *We assume that the residuals $\Phi^{\mathbf{x}}_\sigma$ 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:*

$$\begin{aligned} |C_{\sigma,K}| &> 0 \quad \text{if } \sigma \in K, \\ |C_{\sigma,K}| &= 0 \quad \text{else.} \end{aligned} \tag{36a}$$

and define

$$|C_\sigma| = \sum_{K,\sigma\in K} |C_{\sigma,K}|. \tag{36b}$$

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

$$C_1 \leq \frac{|C_{\sigma,K}|}{|K|} \leq C_2, \tag{37}$$

then there exists $\alpha_1 > 0$ independent of the mesh family such that for any U and V in (25) and (26) with $U_0 = V_0$,

$$\alpha_1 ||U - V|| \leq ||\mathcal{L}^1(U) - \mathcal{L}^1(V)||. \tag{38}$$

Remark 4.5. *The condition is not restrictive because we are solving problems where V^0 is given by the solution at the previous time step. So we are precisely in this setting.*

Before proving this lemma, we have a first one:

Lemma 4.6. *Under the conditions of lemma 4.4, the solution of $\mathcal{L}^1_\sigma(V_1, \dots, V_{r+1}) = |C_\sigma|(A_1, \dots, A_{r+1})^T$ can be explicitely obtained by*

$$\begin{aligned} V_1 &= A_1 + V_0, \\ V_2 &= H_2(V_0 + A_1, A_2), \\ &\vdots \\ V_l &= H_l(V_0 + A_1, A_2, \dots, A_l), \\ &\vdots \\ V_{r+1} &= H_{r+1}(V_0 + A_1, A_2, \dots, A_{r+1}) \end{aligned} \tag{39}$$

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 (37), and because the residuals $\Phi^{\mathbf{x}}_\sigma$ are Lipschitz continuous. □

Proof of lemma 4.4. 1. Clearly, if $A = (A^\sigma)_{\sigma\in\mathcal{S}}$ is a vector of $\mathbb{R}^{|\mathcal{S}|}$, 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 $\mathbb{R}^{|\mathcal{S}|}$, and we are interested in

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

In order to prove (38), it is sufficient to show a similar condition on each of the components of $\mathcal{L}^1$. We first have:

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

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

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

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

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

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

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

Here σ' 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:

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

we have

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

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

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

where C_K depends on the number of vertices of K. Then we conclude by using lemma 4.6 which states that $|U_l - V_l|$ is bounded by $C \sum_{j=1}^{l} |C_\sigma| \, |A_j - B_j|$. Thanks to condition (37), we get the result.

□

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

We write, for any σ, $\mathcal{L}_\sigma^\ell = (\mathcal{L}_{\sigma,0}^\ell, \mathcal{L}_{\sigma,1}^\ell, \ldots, \mathcal{L}_{\sigma,m}^\ell)^T$ 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, $V = (V_1, \ldots, V_{r+1})$ 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

$$\mathcal{L}_{\sigma,p}^1(V) = |C_{\sigma,K}|(V_\sigma^m - V_\sigma^0) + \int_{t_n}^{t_{n,p}} \int_K \mathcal{I}_0(\Phi_\sigma^{\mathbf{x}}(V_0), \ldots, \Phi_\sigma^{\mathbf{x}}(V_{r+1}))ds,$$

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

so that

$$\begin{aligned}
\mathcal{L}_{\sigma,p}^1(V) - \mathcal{L}_{\sigma,p}^2(V) = |C_{\sigma,K}|(V_\sigma^m - V_\sigma^0) - \int_K \psi_\sigma(V^m - V^0) \\
+ \int_{t_n}^{t_{n+1}} \Big(\mathcal{I}_0(\Phi_\sigma^{\mathbf{x}}(V_0), \ldots, \Phi_\sigma^{\mathbf{x}}(V_{r+1})) - \mathcal{I}_l(\Phi_\sigma^{\mathbf{x}}(V_0), \ldots, \Phi_\sigma^{\mathbf{x}}(V_{r+1})) \Big) ds
\end{aligned}$$

We see that to get the estimate (32), a sufficient condition is:

$$\sum_{\sigma \in K} |C_{\sigma,K}|(V_m^\sigma - V_0^\sigma) = \int_K (V_m - V_0) d\mathbf{x} \tag{40}$$

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 4.7. *Under the assumptions of lemma 4.4, 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 4.4. We first have

$$\begin{aligned}\mathcal{L}^1_{\sigma,p}(V) - \mathcal{L}^2_{\sigma,p}(V) = |C_{\sigma,K}|(V_\sigma^m - V_\sigma^0) - \int_K \psi_\sigma(V^m - V^0) \\ + \int_{t_n}^{t_{n+1}} \Big(\mathcal{I}_0(\Phi_\sigma^{\mathbf{x}}(V_0),\dots,\Phi_\sigma^{\mathbf{x}}(V_{r+1})) - \mathcal{I}_l(\Phi_\sigma^{\mathbf{x}}(V_0),\dots,\Phi_\sigma^{\mathbf{x}}(V_{r+1}))\Big)ds\end{aligned}$$

so that

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

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

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

so that

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

Using condition (37) and lemma A.1.

The second term is handled similarly, with the same technique as in the proof of lemma 4.4:

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

Since clearly for any U_j,

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

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 4.4. □

As a consequence, we see that not any basis of test function can be used with this technique: A sufficient condition is that

$$\int_K \varphi_\sigma dx > 0 \tag{41}$$

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 [7, 12] Bézier approximations: if $\lambda_1,\ldots\lambda_{d+1}$ are the barycentric coordinates with respect to the vertices of a simplex, we define, for the multi-index $(i_1,\ldots,i_{d+1})$ with $i_1+\ldots+i_{d+1}=r$

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

We have

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

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

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

Similar properties can be stated for NURBS. In the case of Bézier, the $B_{i_1,\ldots,i_{d+1}}$ 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.

4.4 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 $t_{n,m}$, $m = 0,\ldots,K$, 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 [27], 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.

Another possibility could be to use the basis function of the space defined in [11] for third order and then [20] for higher order since these basis allow exact mass lumping. In this contribution, we have chosen to use Bézier representation since this approach enables better time step constraints. Indeed, the weight defined in [11, 20] can become very small, and hence the quantities C_σ can become very small, in any case much smaller that what is proposed here

4.5 Maximum principle

Here we are interested in the maximum principle. We drop the time subscript, l is the sub-time subscript, u_σ is the solution at the beginning of the computation. In what follows, we use $\mathcal{L}^1$ defined by (29). The $\mathcal{L}^2$ operator, see appendices B and C, writes

$$|K| \sum_{K,\sigma\in K} a^K_{\sigma\sigma'}(V^{(l)}_{p,\sigma'} - V_{\sigma',0}) + \Delta t \sum_{k=1}^{r+1} \theta_{k,r+1}\Big\{ \sum_{\sigma'\in K} c^k_{\sigma\sigma'}(V^{(l)}_{k,\sigma} - V^{(l)}_{k,\sigma'})\Big\}\Big).$$

Here we assume that

- $a^K_{\sigma\sigma} = \gamma^K_\sigma \frac{|K|}{\#K}$ with $\gamma^K_\sigma \in [0,1]$ and $\#K$ is the number of degrees of freedom in K
- and $c^k_{\sigma\sigma'} \geq 0$.

This can be made possible by using the nonlinear RDS schemes described in sections 4 and 5.3, see annex C. We can also assume that $\theta_{k,r+1} \geq 0$, for example by using the strong stability preserving Deferred correction schemes of [23]; they have been designed up to fourth order accuracy. The iterative steps are then:

$$\begin{aligned} |C_\sigma|(V^{l+1}_{p,\sigma} - u_\sigma) &= |C_\sigma|(V^l_{p,\sigma} - u_\sigma) \\ &- \Big(\sum_{K,\sigma\in K} |K|\gamma^K_\sigma \frac{|K|}{\#K}(V^{(l)}_{p,\sigma} - u_\sigma) + \Delta t \sum_{k=1}^{K} \omega_k \Big\{ \sum_{\sigma'\neq\sigma} c^k_{\sigma\sigma'}(V^{(l)}_{k,\sigma} - V^{(l)}_{k,\sigma'})\Big\}\Big) \end{aligned} \tag{42}$$

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

$$\begin{aligned} |C_\sigma|V^{l+1}_{p,\sigma} = \Big(|C_\sigma| - \sum_{K,\sigma\in K} |K|\gamma^K_\sigma \frac{|K|}{\#K} - \Delta t \sum_{k=1}^{K} \omega_k \{ \sum_{\sigma'\neq\sigma} c^k_{\sigma\sigma'}\}\Big) V^l_{p,\sigma} \\ + \Delta t \sum_{k=1}^{K} \omega_k \sum_{\sigma'\neq\sigma} c^k_{\sigma\sigma'} V^{(l)}_{k,\sigma'} \\ + \sum_{K,\sigma\in K} |K|\gamma^K_\sigma \frac{|K|}{\#K} u_\sigma. \end{aligned}$$

Since $|C_\sigma| = \sum_{K,\sigma\in K} \frac{|K|}{\#K}$, $\gamma^K_\sigma \in [0,1]$ 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 the integration rule in time is the trapezoidal one

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

  This gives back the scheme of [26], see appendix D.

- 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 $\mathcal{L}^2$:

$$\int_0^{1/2} \mathcal{I}_2(f)ds = \frac{5}{24}f(0)+\frac{1}{3}f(\frac{1}{2})-\frac{1}{24}f(1)$$
$$\int_0^{1} \mathcal{I}_2(f)ds = \frac{1}{6}f(0)+\frac{4}{6}f(\frac{1}{2})+\frac{1}{6}f(1).$$

  This is nothing more than Simpson formula applied on $[0,1/2]$ where $f(1/4)$ computed using the fact f is polynomial of degree 2 and $[0,1]$. 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 $t=\frac{1}{2}(1+\cos(k\pi/3)$, $k=0,\ldots,3$. We have

$$\int_0^{1/4} \mathcal{I}_3(f)ds = \frac{59}{576}f(0)+\frac{47}{288}f(\frac{1}{4})-\frac{7}{288}f(\frac{1}{2})+\frac{5}{576}f(1)$$
$$\int_0^{3/4} \mathcal{I}_3(f)ds = \frac{3}{64}f(0)+\frac{15}{32}f(\frac{1}{4})+\frac{9}{32}f(\frac{3}{4})-\frac{3}{64}f(1)$$
$$\int_0^{1} \mathcal{I}_3(f)ds = \frac{1}{18}f(0)+\frac{4}{9}f(\frac{1}{4})+\frac{4}{9}f(\frac{3}{4})+\frac{1}{18}f(1)$$

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

The space and time residuals that are tested are: Burman's (20), SUPG (19), and the nonlinear RDS schemes

$$\begin{aligned}\Phi_\sigma(u^h) &= \beta_\sigma(u^h)\Phi_{tot}\\ &+ \sum_{\text{edges of } \partial K} h_e^2||\overline{\nabla_u f}||\int_{\partial K}[\nabla u^h]\cdot[\nabla\varphi_\sigma]d\ell\end{aligned} \tag{43}$$

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

- Choice 1:

$$\beta_\sigma^{PSI}(u^h) = \frac{\left(\frac{\Phi_\sigma^{LxF}}{\Phi_{tot}}\right)^+}{\sum_{\sigma' \in K}\left(\frac{\Phi_{\sigma'}^{LxF}}{\Phi_{tot}}\right)^+}, \tag{44}$$

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

- or Choice 2:

$$\begin{aligned}\beta_\sigma(u^h)\Phi_{tot} &= (1-\theta)\beta_\sigma^{PSI}(u^h)\Phi_{tot} + \theta\Phi_\sigma^{LxF}\\ \theta &= \frac{|\Phi_{tot}|}{\sum_{\sigma' \in K}|\Phi_{LxF}^\sigma|}.\end{aligned} \tag{45}$$

Other choices are possible, for example to blend the Burman or SUPG residual with a Lax Friedrichs one, the blending parameter is defined by θ in choice 2. We nickname the combination (43) and (44) with order k in space and time as BJ-PSI$_k$ and the combination (43) with (45) with order k in space and time as as Bℓ-PSI$_k$.

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 :

$$u_0(x) = \begin{cases} (G(x,\beta,z-\delta) + G(x,\beta,z+\delta) + 4G(x,\beta,z))/6. & \text{if } -0.8 \le x \le -0.6\\ 1 & \text{if } -0.4 \le x \le -0.2\\ 1 - \big|10(x-1)\big| & \text{if } 0 \le x \le 0.2\\ \frac{1}{6}\big(F(x,\alpha,a-\delta) + F(x,\alpha,a+\delta) + 4F(x,\alpha,z)\big) & \text{if } 0.4 \le x \le 0.6\\ 0 & \text{otherwise}\end{cases} \tag{46}$$

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 [21, 19]. 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 Burman residual, to see the effect of increasing the number of iterations. 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

	$\mathbb{B}_2$		$\mathbb{B}_3$	
L^1 error				
$\log_{10} h$	$\log_{10}(err)$	slope	$\log_{10}(err)$	slope
-1.000	-1.7730	-	-2.088	-
-1.301	-3.099	4.406	-4.004	6.35
-1.602	-3.920	2.728	-5.237	4.09
-1.903	-4.740	2.722	-6.425	4.09
-2.204	-5.613	2.900	-7.632	4.09
-2.505	-6.508	2.973	-8.835	4.09
L^2 error				
$\log_{10} h$	$\log_{10}(err)$	slope	$\log_{10}(err)$	slope
-1.000	-1.639	-	-2.310	-
-1.301	-2.851	4.024	-3.498	3.94
-1.602	-3.669	2.719	-4.681	3.93
-1.903	-4.475	2.6760	-5.879	3.98
-2.204	-5.341	2.877	-7.082	3.99
-2.505	-6.234	2.964	-8.286	3.99
L^∞ error				
$\log_{10} h$	$\log_{10}(err)$	slope	$\log_{10}(err)$	slope
-1.000	-1.236	-	-1.617	-
-1.301	-2.309	3.564	-2.843	4.07
-1.602	-3.087	2.585	-3.943	3.65
-1.903	-3.876	2.619	-5.105	3.86
-2.204	-4.730	2.835	-6.299	3.96
-2.505	-5.617	2.948	-7.499	3.96

Table 1: Errors for $u(x,0) = e^{-80(x-0.4)^2}$ and linear advection at $t = 0.25$ with BJ-PSI$_3$ and BJ-PSI$_4$.

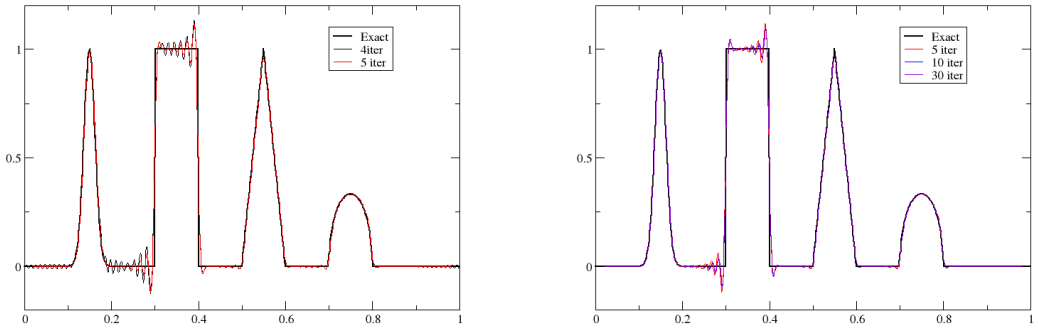
the solutions are aways very well represented, and the resolution improves for the discontinuous part when we increase the number of iterations. Then we have run the same case with non linear schemes (43) with (44). The results are displayed in figure 2. Only 3 iterations are used, the final time is $T = 8$. The results are non oscillatory as expected, and of course more dissipated as in the Burman residual. 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



(a) Burman, 4 and 5 iterations

(b) Burman, 5 and 10, 30 iterations

Figure 1: Results for the convection problem with initial conditions(46), at $T = 8$. The mesh has 200 cells (300 degrees of freedom). The results are obtained with the Burman residual. The mesh has 400 cells (3 degrees of freedom per element).

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

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

where h_K is the length of the smallest edge of K and $\overline{\mathbf{a}}_K$ 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 $\mathbf{B}^2$ 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 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 Burman residuals.

In Figure 4, the same isolines are represented for the three results. We can see that after 10 rotations, the results of the Burman residuals 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 that 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 that 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

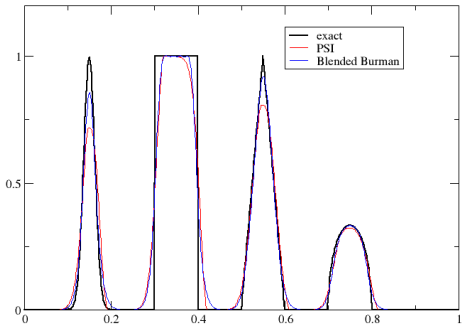


Figure 2: Results for the convection problem with initial conditions(46), at $T = 8$ with quadratic elements. The results are obtained with Bℓ-PSI$_3$ (quadratic elements). Only 3 iterations are made. The mesh has 400 cells (3 degrees of freedom per element).

the SUPG improves a lot, but the cost becomes prohibitive with respect to the Burman residuals 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^∞ norm.

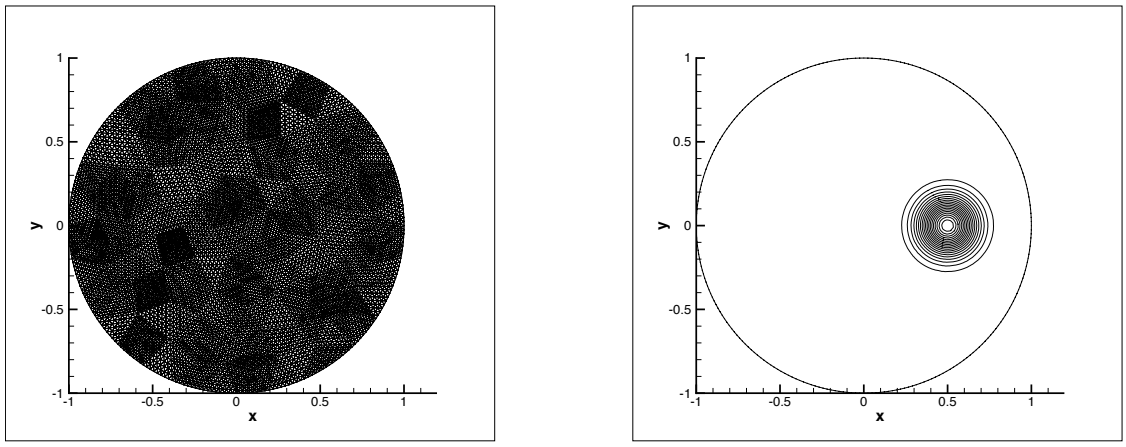
N_{dofs}	$\mathcal{L}^1$	slope	$\mathcal{L}^2$	slope	L^∞	slope
1236	$1.351\ 10^{-1}$	–	$1.335\ 10^{-1}$	–	$5.217\ 10^{-1}$	–
4821	$2.997\ 10^{-2}$	2.21	$4.207\ 10^{-2}$	1.69	$1.967\ 10^{-1}$	1.43
19041	$3.976\ 10^{-3}$	2.94	$7.133\ 10^{-3}$	2.58	$4.149\ 10^{-2}$	2.26
75681	$6.710\ 10^{-4}$	2.57	$1.217\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

Second order

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

Third order

Table 2: Rotation test case with Burman residual with quadratic elements. 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, $h \approx \sqrt{N_{dofs}}$.



(a) Mesh (b) Solution

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

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

The second example is the KPP (Kurganov-Petrov-Petrova) test case, see [22]. The problem is described by

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

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

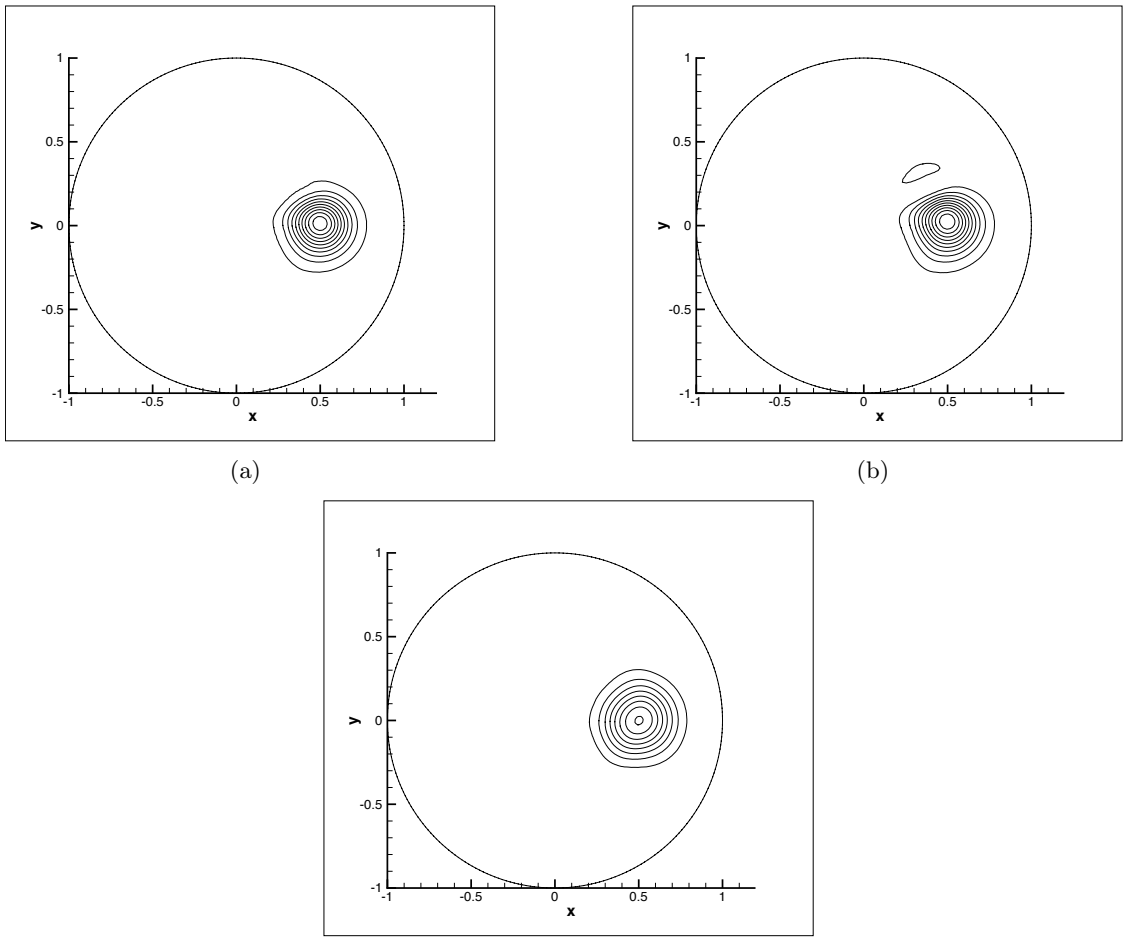
$$u_0(\mathbf{x}) = \begin{cases} \frac{7}{2}\pi & \text{if } ||\mathbf{x}|| < 1 \\ \frac{\pi}{4} & \text{else.} \end{cases}$$

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 $\approx 112.5°$).

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 disapointing, we have considered a spatial approximation using the jump filtering, i.e. the approximation BJ-PSI$_k$ and Bℓ-PSI$_k$. Some results are displayed on Figures 6 and 7. The mesh has been constructed by a mesh generator [15]. 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



(a) (b)

(c)

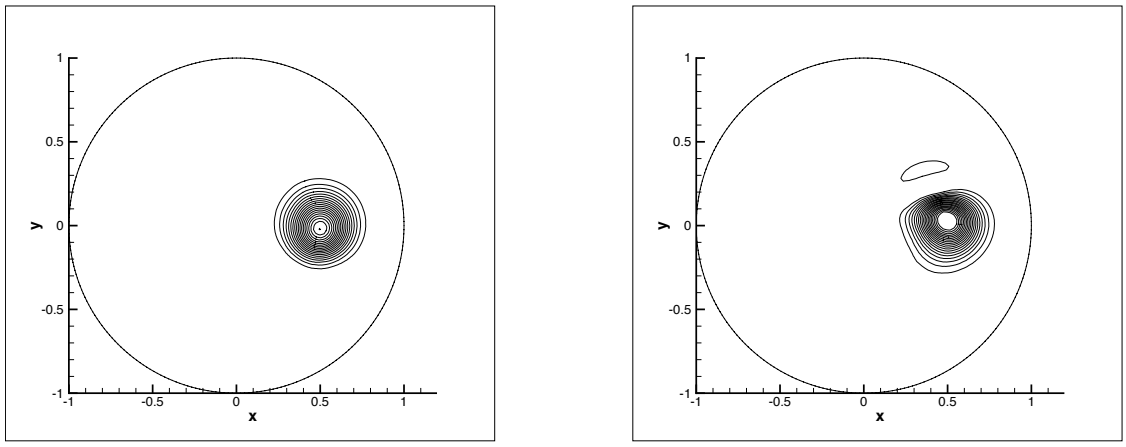
Figure 4: Results for the $\mathbb{P}^1$ approximation: (a) with SUPG, after 1 rotation, (b) with SUPG after 2 rotations, (c) with Burman residual after 10 rotations. The same isolines are represented.

as the one of the linear elements. Both solutions are correct, if one compares to published results, for example [22]. They have been obtained with choice 2 (Bℓ-PSI). If choice 1 (BJ-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 Bℓ-PSI is slightly more dissipative than the BJ-PSI one. We have not represented this solution to save space.

6 Final remarks: About the choice of the temporal scheme

One may wonder why choosing a DeC strategy, instead of more classical strategies. There are several reasons:

1. For some of the spatial discretisation we are interested in, it is not clear if a mass matrix is available. This comes from the fact that the discrete scheme can be reinterpreted in a



(a) (b)

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

discrete variational form, but several different variational forms lead to the same scheme, see for example [2]: how to choose? In addition, are these matrix invertible?This is not clear at all. Because of this it is not possible to use the method of lines.

2. Since the method of line is not possible, one must first discretise the time operator, and has to face the problem of the mass matrix. Possible choices could be the Runge-Kutta methods of the classical Adams or Adams Bashforth methods.

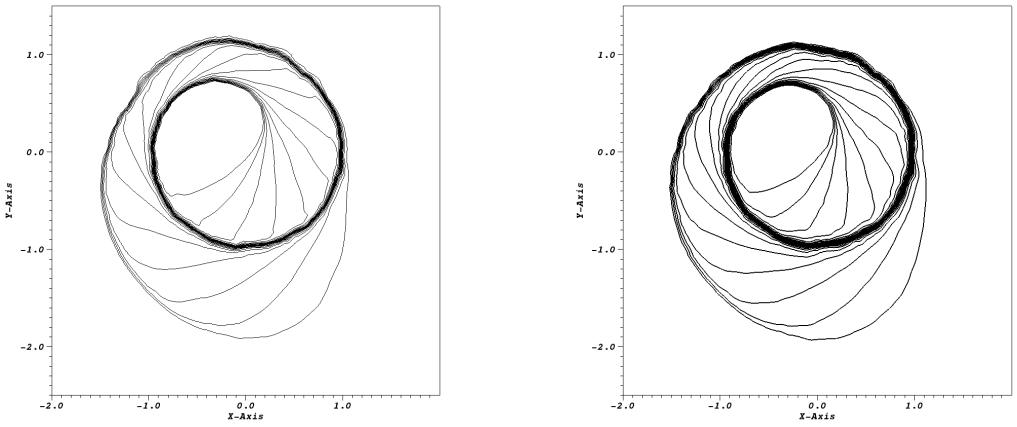
   In the case of Runge Kutta method, one would get a method with no extra cost, in the spirit of [26], provided that the steps that enable to define the RK method are such that the order of accuracy increase by one at every step. This is only true in very special cases This is true for the second order method used in [26], but this is generally wrong. This second order RK method used in [26] can be interpreted as a DeC method, see appendix D. So using RK would necessitate that each step is approximated with full accuracy: the cost would be prohibitive. However, it is possible to use the method of [26] as a building block for high order RK methods suitable with a matrix free approach. We describe one example in the appendix D.

   In the case of the Adams methods, the same kind of iterative method could be used because, from the approximation at the previous time step, one could build a "good" $\mathcal{L}^1$ operator with the order of the Adams method minus one: one integration only would have been needed. However, then there would be no advantage in term of storage, and there would be the problem of the first time steps: we have not tried this approach.

A good compromise is thus the one adopted here.

7 Conclusions and perspectives

In this paper, we have shown how one can get rid of mass matrix for several methods using a globally continuous representation of the solution, including some continuous finite element method. This



(a) $\mathbb{P}^1$ elements, Bℓ-PSI$_2$ (b) B^2 elements, Bℓ-PSI$_3$

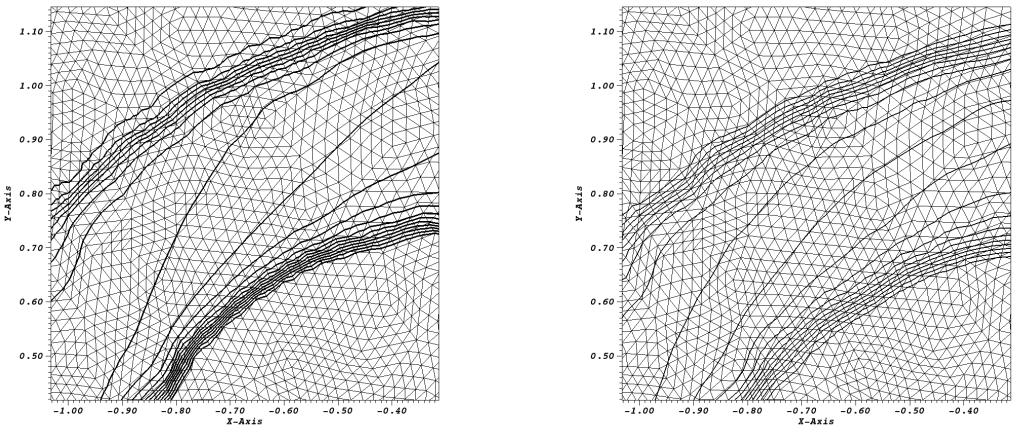
Figure 6: KPP problem, solution at $t = 1$ for the second Bℓ-PSI$_2$ and third order Bℓ-PSI$_3$ schemes.

method relies on a iterative interpretation of the time stepping, in the spirit of Defect Correction method with the use of spatially globally continuous approximation using in each element a polynomial approximation that is spanned on a basis where each basis function has a strictly positive mass. Some analysis indicates that one can get the expected accuracy, this is confirmed by the numerical results obtained on typical linear 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 as the Euler equations. Semi-implicit versions of the schemes can also be constructed in the same spirit. This is work in progress.

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.



(a) $\mathbb{P}^1$ elements, Bℓ-PSI$_2$

(b) B^2 elements, Bℓ-PSI$_3$

Figure 7: KPP problem, solution at $t = 1$ for Bℓ-PSI$_2$ and Bℓ-PSI$_3$. Zoom of the solution, the degrees of freedom are represented.

References

[1] R. Abgrall. Essentially non oscillatory residual distribution schemes for hyperbolic problems. *J. Comput. Phys*, 214(2):773–808, 2006.

[2] R. Abgrall. Residual distribution schemes: current status and future trends. *Computer and Fluids*, 35(7):641–669, 2006.

[3] R. Abgrall. On a class of high order schemes for hyperbolic problems. In *Proceedings of the international Conference of Mathematicians*, volume II, Seoul, 2014.

[4] 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.

[5] 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.

[6] 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.

[7] Rémi Abgrall and Jirka Trefilick. An example of high order residual distribution scheme using non-lagrange elements. *Journal of Scientific Computing*, 45(1-3):64–89, October 2010.

[8] A.Ern and D. di Pietro. *Mathematical aspects of Discontinuous Galerkin Methods*. Mathématiques et Applications. Springer, 2010.

[9] 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.*, 168:464–499, 2001.

[10] E. Burman and P. Hansbo. Edge stabilization for Galerkin approximation of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg*, 193:1437–1453, 2004.

[11] G. Cohen, P. Joly, J.E. Roberts, and N. Tordjman. High order triangular finite element with mass lumping for the wave equation. *SIAM J. Numer. Anal.*, 38:2047–2078, 2001.

[12] 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*, 334:102–124, April 2017.

[13] J. Donea, S. Guilani, and H. Laval. Time-accurate solution of advection-diffusuin problems by finite elements. *Computer Methods in Applied Mechanics and Engineering*, 45(1-3):123–145, 1984.

[14] A. Dutt, L. Greengard, and V. Rokhlin. Spectral deferred correction methods for ordinary differential equations. *BIT Numerical Mathematics*, 40(2):241–266, 2000.

[15] Christophe Geuzaine and Jean-François Remacle. A three-dimensional finite element mesh generator with built-in pre- and post-processing facilities. `http://gmsh.info/`.

[16] E Godlewski and PA Raviart. *Hyperbolic systems of conservation laws*. Ellipses, February 1991.

[17] J.L. Guermond and R. Pasquetti. A correction technique for dispersive effects of mass lumping for transport problems. *Comput. Methods Appl. Mech. Engrg*, 253:186–198, 2013.

[18] 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.

[19] G.S. Jiang and C.W. Shu. Efficient implementation of weighted ENO schemes. *Journal of Computational Physics*, 126:202–228, 1996.

[20] S. Jund and S. Salmon. Arbitrary high-order finite element scheme and high-order mass lumping. *Int. J. Appl. Math. Comput. Sci.*, 17(3):375–393, 2007.

[21] L. Krivodonova. Limiting for high-order discontinuous Galerkin schemes. *Journal of Computational Physics*, 226:879–896, 2007.

[22] 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.

[23] Y. Liu, C.-W. Shu, and M. Zhang. Strong stability preserving property of the defered correction time discretisation. *Journal of Computational Mathematics*, 26(5):633–656, 2008.

[24] R. Löhner, K. Morgan, and O. C. Zienkiewicz. The solution of non-linear hyperbolic equation system by the finite element methods. *International Journal on Numerical methods in Fluids*, 4:1043–1063, 1984.

[25] M.L. Minion. Semi-implicit spectral deferred correction methods for ordinary differential equaions. *Communication in Mathematical Physics*, 1(3):471–500, 2003.

[26] Mario Ricchiuto and Rémi Abgrall. Explicit Runge-Kutta residual-distribution schemes for time dependent problems. *Journal of Computational Physics*, 229(16):5653–5691, 2010.

[27] 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*

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

where C_K is the L^∞ 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.

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^∞ norm of the matrix $(\varphi_{\sigma'})_{\sigma,\sigma'\in K}$. We have:

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

so that

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

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

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

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 Non linear RDS scheme for steady the steady problem (13)

Consider one element K. Since there is no ambiguity, the drop, for the residuals, any reference to K in the following. The total residual is defined by

$$\Phi = \int_{\partial K} \mathbf{f}(u^h) \cdot \mathbf{n}\ d\partial K,$$

and we introduce the Rusanov residuals:

$$\Phi_\sigma^{Rus} = -\int_K \nabla\varphi_\sigma \cdot \mathbf{f}(u^h)\ d\mathbf{x} + \int_{\partial K} \varphi_\sigma(\mathbf{u^h}) \cdot \mathbf{n}\ \mathbf{d}\partial\mathbf{K} + \alpha(\mathbf{u}_\sigma - \bar{\mathbf{u}}),$$

where $\bar{u}$ is the arithmetic average of of the $u'_\sigma s$ on K and α satisfies:

$$\alpha \geq \#K \max_{\sigma,\sigma'\in K} \left| \int_K \varphi_\sigma \nabla\varphi_{\sigma'} \cdot \nabla_u \mathbf{f}(u^h)\ d\mathbf{x} \right|.$$

Here $\#K$ is the number of degrees of freedom in K. This residual can be rewritten as

$$\Phi_\sigma^{Rus} = \sum_{\sigma'\in K} c_{\sigma\sigma'}(u_\sigma - u_{\sigma'})$$

with

$$c_{\sigma\sigma'} = \int_K \varphi_\sigma \nabla\varphi_{\sigma'} \cdot \nabla_u \mathbf{f}(u^h)\, d\mathbf{x} - \frac{\alpha}{\#K}.$$

Under the condition above, $c_{\sigma\sigma'} \geq 0$ and hence we have a maximum principle.

The coefficients β_σ introduced in the relations (22) and (23) are defined by:

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

and can be shown to be always defined, to garanty a local maximum principle for (22) and (23), see [6].

C Some properties of non linear RDS schemes.

This annex is devoted to the justification of some fact stated in section 4.5, namely that the $\mathcal{L}^2$ operator, i.e. for each element and each sub-time step p,

$$\int_K \psi_\sigma(V_p^{(l)} - V_0)dx + \int_K \mathcal{I}_l \Phi_\sigma^{\mathbf{x}}\big(V_0), \ldots, \Phi_\sigma^{\mathbf{x}}(V_{r+1})\big)$$

can write

$$|K| \sum_{\sigma'\in K} a_{\sigma\sigma'}^K (V_{p,\sigma'}^{(l)} - V_{0,\sigma'}) + \Delta t \sum_{k=1}^{r+1} \theta_{k,r+1} \Big\{ \sum_{\sigma'\in K} c_{\sigma\sigma'}^k (V_{k,\sigma}^{(l)} - V_{k,\sigma'}^{(l)}) \Big\} \Big).$$

with

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

In the spirit of section B and [6], we consider the following kind of nonlinear RDS. The Galerkin residuals are defined by

$$\Phi_\sigma^{K,Gal} = \int_{\partial K} \varphi_\sigma \Big(\sum_{k=1}^{r+1} \theta_{k,r+1} \mathbf{f}(V_k^{(l)}) \Big) \cdot \mathbf{n} - \int_K \nabla\varphi_\sigma \Big((\sum_{k=1}^{r+1} \theta_{k,r+1} \mathbf{f}(V_k^{(l)}) \Big), \tag{47}$$

from this one writes a Rusanov residual:

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

with

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

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

$$\Phi_\sigma^{K,\star} = \beta_\sigma^K \Phi_{xt}^K \tag{48}$$

where the total residual Φ^K is defined by

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

and β_σ^K by

$$\beta_\sigma^K = \frac{\max(0, \frac{\Phi_\sigma^{K,Rus}}{\Phi_{xt}^K})}{\sum_{\sigma' \in K} \max(0, \frac{\Phi_\sigma^{K,Rus}}{\Phi_{xt}^K})}. \tag{49}$$

Let us prove now that

$$a_{\sigma\sigma}^K = \gamma_\sigma^K \frac{|K|}{\#K} \text{ with } \gamma_\sigma^K \in [0,1]. \tag{50}$$

and

$$c_{\sigma\sigma'}^k \geq 0. \tag{51}$$

Proof. We note that

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

with

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

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

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

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 (50).

The second one (51) comes from the very definition of β_σ^K. □

Remark C.1. *In many practical applications, the residual that is considered is not* (48) *but*

$$\Phi_\sigma^{K,\star} = \beta_\sigma^K \Phi_{xt}^K + \sum_{edges\ of\ K} h_K^2 \Gamma \int_e \left[\nabla \varphi_\sigma \right] \left[\nabla \left(\sum_{k=1}^{r+1} \theta_{k,r+1} V_k^{(l)} \right) \right] \tag{52}$$

with β_σ^K defined as (49)*. Then* (50) *is still true because the term*

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

does not contain any time increment.

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

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

with now

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

and we consider

$$\Phi_\sigma^{K,\star} = (1-\ell) \Big\{ \Phi_\sigma^{K,Gal} + \sum_{edges\ of\ K} h_K^2 \Gamma \int_e \Big[\nabla \varphi_\sigma \Big] \Big[\nabla \Big(\sum_{k=1}^{r+1} \theta_{k,r+1} V_k^{(l)} \Big) \Big] \Big\} + \ell \Phi_\sigma^{K,Rus} \tag{53}$$

with

$$\ell = \frac{|\Phi^K|}{\sum_{\sigma \in K} |\Phi_\sigma^{K,Rus}|}. \tag{54}$$

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

D Example of suitable Runge Kutta method

Consider the problem

$$\frac{dy}{dt} = f(y)$$

wit $y_0 = y(0)$. The Runge-Kutta method of [26] is:

$$\begin{aligned} y^{(0)} &= u^n \\ y^{(1)} &= y^{(0)} + \Delta t f(y^{(0)}) \\ y^{(2)} &= y^{(1)} + \frac{\Delta t}{2} \Big(f(y^{(0)}) + f(y^{(1)}) \Big) \end{aligned} \tag{55}$$

We first show that this is indeed a DeC method.

Defining

$$\mathcal{L}^2(Y) = Y - Y_0 + \frac{\Delta t}{2} \big(f(Y_0) + f(Y) \big)$$

and

$$\mathcal{L}^{(1)}(Y) = Y - Y_0 + \Delta t f(Y_0),$$

The first iteration leads to

$$y^{(1)} - y^{(0)} + \Delta t f(y^{(0)}) = \Big(y^{(0)} - y^{(0)} + \Delta t f(y^{(0)}) \Big) - \Big(y^{(0)} - y^{(0)} + \frac{\Delta t}{2} \Big(f(y^{(0)}) + f(y^{(0)}) \Big)$$

i.e.

$$y^{(1)} = y^{(0)} + \Delta t f(y^{(0)}).$$

The second iteration gives:

$$y^{(2)} - y^{(0)} + \Delta t f(y^{(0)}) = \Big(y^{(1)} - y^{(0)} + \Delta t f(y^{(0)}) \Big) - \Big(y^{(1)} - y^{(0)} + \frac{\Delta t}{2} \Big(f(y^{(0)}) + f(y^{(1)}) \Big)$$

i.e.

$$y^{(2)} = y^{(1)} + \frac{\Delta t}{2} \Big(f(y^{(0)}) + f(y^{(1)}) \Big).$$

From this, let us construct an example of RK scheme which is third order, using this method as a building block. We know the exact solution satisfies:

$$y(t_{n+1}) = y(t_n)) + \int_{t_n}^{t_{n+1}} f(y(s)) ds.$$

Using Simpson formula, an approximation is:

$$y_{n+1} = y_n + \frac{\Delta}{6} \Big(f(y_n) + 4 f(y_{n+1/2}) + f(y_{n+1}) \Big).$$

To get a third odred approximation of this third order approximation of the exact solution, we need second order approximations of $y_{n+1/2} \approx y(t_n + \frac{\Delta t}{2})$ and $y_{n+1} \approx y(t_{n+1})$.

To do so, we can use the method (55) to extimate y_{n+1}. And using the intermediate steps at $t_n + \frac{\Delta t}{4}$ and $t_n + \frac{\Delta t}{2} = t_{n+1/2}$, with (55) we get a second order approximation of $y_{n+1/2}$.

Note we can save one iteration (or one step) by first evaluating $y_{n+1/2}$ with (55) and using the second step of (55) to get a second order approximation of y_{n+1}.

We have used this method for convextion problems, and we get similar results as with the method presented here, which is more systematic.

This can be extended as follows: one first start by a quadrature formula of order k of $\int_{t_n}^{t_{n+1}} f(y(s))$ and we iteratively get approximations of order $k-1$ of the quadrature points by induction.