

A new efficient explicit Deferred Correction framework: analysis and applications to hyperbolic PDEs and adaptivity Supplementary Material

L. Micalizzi* and D. Torlo†

October 6, 2022

Introduction

In this supplementary material, we show the proofs and the details that were too lengthy to be put in the principal manuscript. We show the proof of the Deferred Correction procedure in a general framework in section 1. In section 2, we provide the proofs of the accuracy and of the properties of the operators \mathcal{L}_Δ^2 and \mathcal{L}_Δ^1 of the bDeC method in the context of ODEs, and we show how the sDeC method can be seen as a perturbation of the bDeC. In section 3, we prove the properties of the operators \mathcal{L}_Δ^2 and \mathcal{L}_Δ^1 of the bDeC formulation for the continuous Galerkin (CG) finite element framework and we investigate the issues experienced in many works with such formulation. Finally, in section 4, we show how to find the analytical solution to the ODE modeling a monodimensional vibrating system.

For each section, we recall the basic notions of the main document needed for the discussion, in order to make this document as much self-contained as possible, and sometimes deepened, in order to increase the understandability.

1 DeC in the abstract framework

Assume that we have two operators, depending on a parameter Δ , between two normed vector spaces

$$\mathcal{L}_\Delta^1, \mathcal{L}_\Delta^2 : X \longrightarrow Y \tag{1}$$

then the following theorem holds.

Theorem 1.1 (Deferred Correction accuracy). *Let the following hypotheses hold*

1. **Existence of a unique solution to \mathcal{L}_Δ^2**
 $\exists! \mathbf{u}_\Delta \in X$ solution of \mathcal{L}_Δ^2 such that $\mathcal{L}_\Delta^2(\mathbf{u}_\Delta) = \mathbf{0}_Y$;

*Affiliation: Institute of Mathematics, University of Zurich, Winterthurerstrasse 190, Zurich, 8057, Switzerland.
Email: lorenzo.micalizzi@math.uzh.ch.

†Affiliation: SISSA mathLab, SISSA, via Bonomea 265, Trieste, 34136, Italy. Email: davide.torlo@sissa.it.

2. **Coercivity-like property of \mathcal{L}_Δ^1**
 $\exists \alpha_1 \geq 0$ independent of Δ s.t.

$$\|\mathcal{L}_\Delta^1(\underline{\mathbf{v}}) - \mathcal{L}_\Delta^1(\underline{\mathbf{w}})\|_Y \geq \alpha_1 \|\underline{\mathbf{v}} - \underline{\mathbf{w}}\|_X, \quad \forall \underline{\mathbf{v}}, \underline{\mathbf{w}} \in X; \quad (2)$$

3. **Lipschitz-continuity-like property of $\mathcal{L}_\Delta^1 - \mathcal{L}_\Delta^2$**
 $\exists \alpha_2 \geq 0$ independent of Δ s.t.

$$\|(\mathcal{L}_\Delta^1(\underline{\mathbf{v}}) - \mathcal{L}_\Delta^2(\underline{\mathbf{v}})) - (\mathcal{L}_\Delta^1(\underline{\mathbf{w}}) - \mathcal{L}_\Delta^2(\underline{\mathbf{w}}))\|_Y \leq \alpha_2 \Delta \|\underline{\mathbf{v}} - \underline{\mathbf{w}}\|_X, \quad \forall \underline{\mathbf{v}}, \underline{\mathbf{w}} \in X. \quad (3)$$

Then, if we iteratively define $\underline{\mathbf{u}}^{(p)}$ as the solution of

$$\mathcal{L}_\Delta^1(\underline{\mathbf{u}}^{(p)}) = \mathcal{L}_\Delta^1(\underline{\mathbf{u}}^{(p-1)}) - \mathcal{L}_\Delta^2(\underline{\mathbf{u}}^{(p-1)}), \quad p = 1, \dots, P, \quad (4)$$

we have that

$$\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}_\Delta\|_X \leq \left(\Delta \frac{\alpha_2}{\alpha_1}\right)^P \|\underline{\mathbf{u}}^{(0)} - \underline{\mathbf{u}}_\Delta\|_X. \quad (5)$$

Proof. By using the coercivity-like property of \mathcal{L}_Δ^1 and the definition of $\mathcal{L}_\Delta^1(\underline{\mathbf{u}}^{(p)})$ in (4), we have

$$\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}_\Delta\|_X \leq \frac{1}{\alpha_1} \|\mathcal{L}_\Delta^1(\underline{\mathbf{u}}^{(P)}) - \mathcal{L}_\Delta^1(\underline{\mathbf{u}}_\Delta)\|_Y = \frac{1}{\alpha_1} \|\mathcal{L}_\Delta^1(\underline{\mathbf{u}}^{(P-1)}) - \mathcal{L}_\Delta^2(\underline{\mathbf{u}}^{(P-1)}) - \mathcal{L}_\Delta^1(\underline{\mathbf{u}}_\Delta)\|_Y. \quad (6)$$

Since $\underline{\mathbf{u}}_\Delta$ is the solution of \mathcal{L}_Δ^2 , we have that $\mathcal{L}_\Delta^2(\underline{\mathbf{u}}_\Delta) = \mathbf{0}_Y$ and we can add it inside the norm on the right hand side of the equality in (6) and we get

$$\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}_\Delta\|_X \leq \frac{1}{\alpha_1} \left\| \left[\mathcal{L}_\Delta^1(\underline{\mathbf{u}}^{(P-1)}) - \mathcal{L}_\Delta^2(\underline{\mathbf{u}}^{(P-1)}) \right] - \left[\mathcal{L}_\Delta^1(\underline{\mathbf{u}}_\Delta) - \mathcal{L}_\Delta^2(\underline{\mathbf{u}}_\Delta) \right] \right\|_Y. \quad (7)$$

Now, by applying the Lipschitz-continuity-like property we get

$$\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}_\Delta\|_X \leq \Delta \frac{\alpha_2}{\alpha_1} \|\underline{\mathbf{u}}^{(P-1)} - \underline{\mathbf{u}}_\Delta\|_X. \quad (8)$$

By repeating these calculations recursively we get the thesis. \square

2 The Deferred Correction for systems of ODEs

We will focus on the numerical solution of the general Cauchy problem

$$\begin{cases} \frac{d}{dt} \mathbf{u}(t) = \mathbf{G}(t, \mathbf{u}(t)), & t \in [0, T], \\ \mathbf{u}(0) = \mathbf{z}, \end{cases} \quad (9)$$

with $\mathbf{u} : \mathbb{R}_0^+ \rightarrow \mathbb{R}^Q$, $\mathbf{z} \in \mathbb{R}^Q$ and $\mathbf{G} : \mathbb{R}_0^+ \times \mathbb{R}^Q \rightarrow \mathbb{R}^Q$ a continuous map Lipschitz continuous with respect to \mathbf{u} uniformly with respect to t with a Lipschitz constant L . This ensures the existence of a unique solution for the system of ODEs (9).

We will assume here a classical one-step method setting: we discretize the time domain $[0, T]$ by introducing $N + 1$ time nodes t_n , which are such that $0 = t_0 < t_1 < \dots < t_N = T$ and therefore inducing N intervals $[t_n, t_{n+1}]$, we denote by \mathbf{u}_n an approximation of the exact solution $\mathbf{u}(t_n)$ at the time t_n and we look for a recipe to compute \mathbf{u}_{n+1} by knowing \mathbf{u}_n for each $n = 0, 1, \dots, N - 1$. We will focus on the generic time interval $[t_n, t_{n+1}]$ with $\Delta t = t_{n+1} - t_n$ and, as in the context of a general consistency analysis, we will assume $\mathbf{u}_n = \mathbf{u}(t_n)$.

2.1 bDeC

In the general time step $[t_n, t_n + \Delta t]$ we introduce $M + 1$ subtimenodes t^0, \dots, t^M such that $t_n = t^0 < t^1 < \dots < t^M = t_n + \Delta t$, which are assumed here to be equispaced. We will refer to $\mathbf{u}(t^m)$ as the exact solution in the node t^m and to \mathbf{u}^m as the approximation of the solution in the same node. Just for the first node, we set $\mathbf{u}^0 := \mathbf{u}_n$ and, in the accuracy study, we will consider it to be exact, i.e., $\mathbf{u}^0 = \mathbf{u}(t^0) = \mathbf{u}(t_n) = \mathbf{u}_n$.

2.1.1 Definition of \mathcal{L}_Δ^2

An exact integration of the system of ODEs over $[t^0, t^m]$ would result in

$$\mathbf{u}(t^m) - \mathbf{u}^0 - \int_{t^0}^{t^m} \mathbf{G}(t, \mathbf{u}(t)) dt = \mathbf{0}, \quad \forall m = 1, \dots, M, \quad (10)$$

from which we would have the exact solution $\mathbf{u}(t^m)$.

Unfortunately, we cannot perform in general the exact integration and we need to make some approximations. We replace $\mathbf{G}(t, \mathbf{u}(t))$ by the Lagrange interpolating polynomial of degree M associated to the $M + 1$ nodes t^m with $m = 0, 1, \dots, M$, getting

$$\mathbf{u}^m - \mathbf{u}^0 - \int_{t^0}^{t^m} \sum_{\ell=0}^M \mathbf{G}(t^\ell, \mathbf{u}(t^\ell)) \psi^\ell(t) dt = \mathbf{0}, \quad \forall m = 1, \dots, M. \quad (11)$$

Moving the finite sum and the vectors $\mathbf{G}(t^\ell, \mathbf{u}(t^\ell))$ outside of the integral, (11) can be recast as

$$\mathbf{u}^m - \mathbf{u}^0 - \Delta t \sum_{\ell=0}^M \theta_\ell^m \mathbf{G}(t^\ell, \mathbf{u}(t^\ell)) = \mathbf{0}, \quad \forall m = 1, \dots, M, \quad (12)$$

where the coefficients θ_ℓ^m are the normalized integrals of the Lagrange basis functions and do not depend on Δt .

Proposition 2.1. *\mathbf{u}^m satisfying (12) is an $(M + 1)$ -order accurate approximation of $\mathbf{u}(t^m)$.*

Proof. For the proof, we will focus on the original equivalent formulation (11). Let us compute $\mathbf{u}(t^m) - \mathbf{u}^m$ with \mathbf{u}^m got by (11). From (10), (11) and the M -order accuracy on the approximation of $\mathbf{G}(t, \mathbf{u}(t))$ due to the interpolation with Lagrange polynomials of degree M we have

$$\begin{aligned} \mathbf{u}(t^m) - \mathbf{u}^m &= \mathbf{u}^0 + \int_{t^0}^{t^m} \mathbf{G}(t, \mathbf{u}(t)) dt - \mathbf{u}^0 - \int_{t^0}^{t^m} \sum_{\ell=0}^M \mathbf{G}(t^\ell, \mathbf{u}(t^\ell)) \psi^\ell(t) dt \\ &= \int_{t^0}^{t^m} \left[\mathbf{G}(t, \mathbf{u}(t)) - \sum_{\ell=0}^M \mathbf{G}(t^\ell, \mathbf{u}(t^\ell)) \psi^\ell(t) \right] dt \\ &= \int_{t^0}^{t^m} O(\Delta t^{M+1}) dt = O(\Delta t^{M+2}). \end{aligned} \quad (13)$$

□

Despite this result, the previous formula cannot be used in practice because the exact solution $\mathbf{u}(t^\ell)$ in the nodes t^ℓ with $\ell = 1, \dots, M$ is not available.

We use the approximated values \mathbf{u}^ℓ in place of them, thus getting the following implicit formulation

$$\mathbf{u}^m - \mathbf{u}^0 - \Delta t \sum_{\ell=0}^M \theta_\ell^m \mathbf{G}(t^\ell, \mathbf{u}^\ell) = \mathbf{0} \quad \forall m = 1, \dots, M, \quad (14)$$

which leads to the definition of our \mathcal{L}_Δ^2 operator

$$\mathcal{L}_\Delta^2(\underline{\mathbf{u}}) = \begin{pmatrix} \mathbf{u}^1 - \mathbf{u}^0 - \Delta t \sum_{\ell=0}^M \theta_\ell^1 \mathbf{G}(t^\ell, \mathbf{u}^\ell) \\ \vdots \\ \mathbf{u}^m - \mathbf{u}^0 - \Delta t \sum_{\ell=0}^M \theta_\ell^m \mathbf{G}(t^\ell, \mathbf{u}^\ell) \\ \vdots \\ \mathbf{u}^M - \mathbf{u}^0 - \Delta t \sum_{\ell=0}^M \theta_\ell^M \mathbf{G}(t^\ell, \mathbf{u}^\ell) \end{pmatrix} \quad \text{with } \underline{\mathbf{u}} = \begin{pmatrix} \mathbf{u}^1 \\ \vdots \\ \mathbf{u}^m \\ \vdots \\ \mathbf{u}^M \end{pmatrix}. \quad (15)$$

Proposition 2.2. *Let \mathbf{u}^m be the m -th component of the solution of $\mathcal{L}_\Delta^2(\underline{\mathbf{u}}) = \mathbf{0}$. Then, \mathbf{u}^m is an $(M + 1)$ -order accurate approximation of $\mathbf{u}(t^m)$.*

Proof. Let us consider the following operator $\mathcal{J} : \mathbb{R}^{(M \times Q)} \rightarrow \mathbb{R}^{(M \times Q)}$ defined as

$$\underline{\mathbf{y}} = \mathcal{J}(\underline{\mathbf{u}}) = \begin{pmatrix} \mathbf{u}^0 + \Delta t \sum_{\ell=0}^M \theta_\ell^1 \mathbf{G}(t^\ell, \mathbf{u}^\ell) \\ \vdots \\ \mathbf{u}^0 + \Delta t \sum_{\ell=0}^M \theta_\ell^m \mathbf{G}(t^\ell, \mathbf{u}^\ell) \\ \vdots \\ \mathbf{u}^0 + \Delta t \sum_{\ell=0}^M \theta_\ell^M \mathbf{G}(t^\ell, \mathbf{u}^\ell) \end{pmatrix} \quad \text{with } \underline{\mathbf{y}} = \begin{pmatrix} \mathbf{y}^1 \\ \vdots \\ \mathbf{y}^m \\ \vdots \\ \mathbf{y}^M \end{pmatrix}. \quad (16)$$

Again, we remark that \mathbf{u}^0 , the vector corresponding to the initial subtimenode, is always fixed. The proof consists of two parts. We will first show that, for Δt small enough, \mathcal{J} is a contraction over $\mathbb{R}^{(M \times Q)}$, which is a finite dimensional space (and so complete with respect to the distance induced by any norm). This will ensure, thanks to the Banach fixed-point theorem, that there exists a fixed point $\tilde{\underline{\mathbf{u}}}$ such that $\tilde{\underline{\mathbf{u}}} = \mathcal{J}(\tilde{\underline{\mathbf{u}}})$ and that it is unique. It is very easy to see that this fixed point is the (unique) solution to the operator \mathcal{L}_Δ^2 . Then, by iteratively applying the operator, we will generate a sequence of vectors converging to this fixed point and we will show that this limit is an $(M + 1)$ -order accurate approximation of the exact solution to the system of ODEs.

Let us first prove that \mathcal{J} is a contraction for Δt small enough. We recall that θ_ℓ^m are constant coefficients independent on Δt and bounded by $C_\theta = \max |\theta_\ell^m|$ and that $\mathbf{G}(t, \mathbf{u})$ is Lipschitz-continuous with respect to \mathbf{u} uniformly with respect to t with constant L . Now, using the triangular

inequality, we have

$$\begin{aligned}
\|\mathcal{J}(\underline{\mathbf{v}}) - \mathcal{J}(\underline{\mathbf{w}})\|_\infty &= \Delta t \left\| \sum_{\ell=0}^M \begin{pmatrix} \theta_\ell^M [\mathbf{G}(t^\ell, \mathbf{v}^\ell) - \mathbf{G}(t^\ell, \mathbf{w}^\ell)] \\ \vdots \\ \theta_\ell^m [\mathbf{G}(t^\ell, \mathbf{v}^\ell) - \mathbf{G}(t^\ell, \mathbf{w}^\ell)] \\ \vdots \\ \theta_\ell^1 [\mathbf{G}(t^\ell, \mathbf{v}^\ell) - \mathbf{G}(t^\ell, \mathbf{w}^\ell)] \end{pmatrix} \right\|_\infty \\
&\leq \Delta t C_\theta \sum_{\ell=0}^M \|\mathbf{G}(t^\ell, \mathbf{v}^\ell) - \mathbf{G}(t^\ell, \mathbf{w}^\ell)\|_{\infty, Q} \\
&\leq \Delta t C_\theta \sum_{\ell=0}^M L \|\mathbf{v}^\ell - \mathbf{w}^\ell\|_{\infty, Q} \leq \Delta t C_\theta L M \|\underline{\mathbf{v}} - \underline{\mathbf{w}}\|_\infty.
\end{aligned} \tag{17}$$

The last inequality follows from the fact that $\underline{\mathbf{v}} - \underline{\mathbf{w}}$ contains as components all the vectors $\mathbf{v}^\ell - \mathbf{w}^\ell$ for all $\ell = 1, \dots, M$ and from the fact that $\mathbf{v}^0 = \mathbf{w}^0 = \mathbf{u}^0$ and so

$$\|\mathbf{v}^\ell - \mathbf{w}^\ell\|_{\infty, Q} \leq \|\underline{\mathbf{v}} - \underline{\mathbf{w}}\|_\infty, \quad \forall \ell = 1, \dots, M, \tag{18}$$

where $\|\cdot\|_{\infty, Q}$ is the infinity norm over \mathbb{R}^Q , while $\|\cdot\|_\infty$ is the infinity norm over $\mathbb{R}^{M \times Q}$. For $\Delta t < \frac{1}{C_\theta L M}$, we have

$$\|\mathcal{J}(\underline{\mathbf{v}}) - \mathcal{J}(\underline{\mathbf{w}})\|_\infty < \delta \|\underline{\mathbf{v}} - \underline{\mathbf{w}}\|_\infty \tag{19}$$

with $\delta < 1$ and so \mathcal{J} is a contraction. As anticipated, there exists a unique fixed point $\underline{\hat{\mathbf{u}}}$, solution of \mathcal{L}_Δ^2 .

For the second part, we will prove the accuracy of the iteration of the fixed point procedure. We consider the sequence $\{\underline{\mathbf{y}}^{(k)}\}_{k \in \mathbb{N}}$ given by the following recursive definition

$$\underline{\mathbf{y}}^{(k)} = \mathcal{J}(\underline{\mathbf{y}}^{(k-1)}) \tag{20}$$

with its general element being

$$\underline{\mathbf{y}}^{(k)} = \begin{pmatrix} \mathbf{y}^{1,(k)} \\ \vdots \\ \mathbf{y}^{m,(k)} \\ \vdots \\ \mathbf{y}^{M,(k)} \end{pmatrix}, \text{ with } \underline{\mathbf{y}}^{(0)} = \begin{pmatrix} \mathbf{y}^{1,(0)} \\ \vdots \\ \mathbf{y}^{m,(0)} \\ \vdots \\ \mathbf{y}^{M,(0)} \end{pmatrix} = \begin{pmatrix} \mathbf{u}^0 \\ \vdots \\ \mathbf{u}^0 \\ \vdots \\ \mathbf{u}^0 \end{pmatrix}. \tag{21}$$

The general component $\mathbf{y}^{m,(k)}$ of $\underline{\mathbf{y}}^{(k)}$ is a Q -dimensional vector. The first index m is referred to the subtimenode, the second is the index of the sequence. In order to have a more compact notation, we will not write $\mathbf{G}(t^0, \mathbf{u}^0)$ as a separate term but we set $\mathbf{y}^{0,(k)} = \mathbf{u}^0 \forall k \geq 0$, because the value of the solution at the first subtimenode is known. From theory, we know that this sequence converges to the fixed point of \mathcal{J} and so to the solution of the operator \mathcal{L}_Δ^2 .

Let us prove by induction on k that for all $m = 1, \dots, M$, we have

$$\mathbf{y}^{m,(k)} = \mathbf{u}(t^m) + O(\Delta t^{\min(k+1, M+2)}). \tag{22}$$

The base case, for $k = 0$, is clearly true as a simple Taylor expansion gives

$$\mathbf{u}(t^m) = \mathbf{u}(t^0) + \Delta t \mathbf{G}(t^0, \mathbf{u}(t^0))(t^m - t^0) + O(\Delta t^2) = \mathbf{y}^{m,(0)} + O(\Delta t), \quad (23)$$

reminding that $\frac{d}{dt}\mathbf{u}(t) = \mathbf{G}(t, \mathbf{u}(t))$.

For the induction step, we assume that $\mathbf{y}^{m,(k)} = \mathbf{u}(t^m) + O(\Delta t^{\min(k+1, M+2)})$ and we will prove that $\mathbf{y}^{m,(k+1)} = \mathbf{u}(t^m) + O(\Delta t^{\min(k+2, M+2)})$. By exploiting the Lipschitz-continuity of \mathbf{G} , we have that

$$\begin{aligned} \mathbf{G}(t^\ell, \mathbf{u}(t^\ell)) &= \mathbf{G}(t^\ell, \mathbf{y}^{\ell,(k)}) + \nabla_{\mathbf{u}} \mathbf{G}(t^\ell, \mathbf{y}^{\ell,(k)})(\mathbf{u}(t^\ell) - \mathbf{y}^{\ell,(k)}) + O\left(\left\|\mathbf{u}(t^\ell) - \mathbf{y}^{\ell,(k)}\right\|_{\infty, Q}^2\right) \\ &= \mathbf{G}(t^\ell, \mathbf{y}^{\ell,(k)}) + O(\Delta t^{\min(k+1, M+2)}), \end{aligned} \quad (24)$$

where $\nabla_{\mathbf{u}} \mathbf{G}(t^\ell, \mathbf{y}^{\ell,(k)})$ is bounded in some norm by L . We are then able to prove that

$$\begin{aligned} \mathbf{y}^{m,(k+1)} &= \mathbf{u}(t^0) + \Delta t \sum_{\ell=0}^M \theta_\ell^m \mathbf{G}(t^\ell, \mathbf{y}^{\ell,(k)}) \\ &= \mathbf{u}(t^0) + \Delta t \sum_{\ell=0}^M \theta_\ell^m \mathbf{G}(t^\ell, \mathbf{u}(t^\ell)) + O(\Delta t^{1+\min(k+1, M+2)}). \end{aligned} \quad (25)$$

Now, thanks to the $(M+1)$ -order accuracy of (12), we have that

$$\begin{aligned} \mathbf{y}^{m,(k+1)} &= \mathbf{u}(t^0) + \Delta t \sum_{\ell=0}^M \theta_\ell^m \mathbf{G}(t^\ell, \mathbf{u}(t^\ell)) + O(\Delta t^{1+\min(k+1, M+2)}) \\ &= \mathbf{u}(t^m) + O(\Delta t^{M+2}) + O(\Delta t^{1+\min(k+1, M+2)}) = \mathbf{u}(t^m) + O(\Delta t^{\min(k+2, M+2)}). \end{aligned} \quad (26)$$

Hence, for $k > M$ the components $\mathbf{y}^{(k),m}$ are an $(M+1)$ accurate solution of $\mathbf{u}(t^m)$ and their limit for $k \rightarrow \infty$, i.e., the solutions of \mathcal{L}_Δ^2 , is as well an $(M+1)$ approximation of the exact solution. \square

2.1.2 Definition of \mathcal{L}_Δ^1

If we apply the Euler method to get the approximate solution \mathbf{u}^m in the node t^m we have

$$\mathbf{u}^m - \mathbf{u}^0 - \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) = \mathbf{0}, \quad (27)$$

where $\beta^m = \frac{t^m - t^0}{\Delta t}$.

Proposition 2.3. *Let \mathbf{u}^m be the solution of (27), then \mathbf{u}^m is first order accurate, i.e., $\mathbf{u}(t^m) - \mathbf{u}^m = O(\Delta t^2)$.*

Proof. We consider the difference between the exact solution $\mathbf{u}(t^m)$ to our ODEs system and \mathbf{u}^m got from (27). Through a first order Taylor expansion of $\mathbf{u}(t)$ and from the fact that $\frac{d}{dt}\mathbf{u}(t) = \mathbf{G}(t, \mathbf{u}(t))$, we have

$$\mathbf{u}(t^m) - \mathbf{u}^m = \mathbf{u}^0 + \mathbf{G}(t^0, \mathbf{u}^0)(t^m - t^0) + O(\Delta t^2) - \mathbf{u}^0 - \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) = O(\Delta t^2), \quad (28)$$

because $\mathbf{u}^0 = \mathbf{u}(t^0) = \mathbf{u}(t_n) = \mathbf{u}_n$ and $\beta^m = \frac{t^m - t^0}{\Delta t}$. \square

Directly from (27), we get our explicit, low order operator $\mathcal{L}_\Delta^1 : \mathbb{R}^{(M \times Q)} \rightarrow \mathbb{R}^{(M \times Q)}$ defined as

$$\mathcal{L}_\Delta^1(\underline{\mathbf{u}}) = \begin{pmatrix} \mathbf{u}^1 - \mathbf{u}^0 - \Delta t \beta^1 \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{u}^m - \mathbf{u}^0 - \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{u}^M - \mathbf{u}^0 - \Delta t \beta^M \mathbf{G}(t^0, \mathbf{u}^0) \end{pmatrix} \text{ with } \underline{\mathbf{u}} = \begin{pmatrix} \mathbf{u}^1 \\ \vdots \\ \mathbf{u}^m \\ \vdots \\ \mathbf{u}^M \end{pmatrix}. \quad (29)$$

2.1.3 Proof of the properties of \mathcal{L}_Δ^1 and \mathcal{L}_Δ^2

We equip $X = Y = \mathbb{R}^{(M \times Q)}$ with the infinity norm $\|\cdot\|_\infty$ and we recall here the hypotheses that are needed to apply the Deferred Correction method from the abstract formulation but characterizing them to our case.

- i) **Existence of a solution to \mathcal{L}_Δ^2**
 $\exists! \underline{\mathbf{u}}_\Delta \in \mathbb{R}^{(M \times Q)}$ solution of \mathcal{L}_Δ^2 , i.e. such that $\mathcal{L}_\Delta^2(\underline{\mathbf{u}}_\Delta) = \mathbf{0}$;
- ii) **Coercivity-like property of \mathcal{L}_Δ^1**
 $\exists \alpha_1 \geq 0$ independent of Δt s.t.

$$\|\mathcal{L}_\Delta^1(\underline{\mathbf{v}}) - \mathcal{L}_\Delta^1(\underline{\mathbf{w}})\|_\infty \geq \alpha_1 \|\underline{\mathbf{v}} - \underline{\mathbf{w}}\|_\infty, \quad \forall \underline{\mathbf{v}}, \underline{\mathbf{w}} \in \mathbb{R}^{(M \times Q)}; \quad (30)$$

- iii) **Lipschitz-continuity-like condition of $\mathcal{L}_\Delta^1 - \mathcal{L}_\Delta^2$**
 $\exists \alpha_2 \geq 0$ independent of Δt s.t.

$$\|[\mathcal{L}_\Delta^1(\underline{\mathbf{v}}) - \mathcal{L}_\Delta^2(\underline{\mathbf{v}})] - [\mathcal{L}_\Delta^1(\underline{\mathbf{w}}) - \mathcal{L}_\Delta^2(\underline{\mathbf{w}})]\|_\infty \leq \alpha_2 \Delta t \|\underline{\mathbf{v}} - \underline{\mathbf{w}}\|_\infty, \quad \forall \underline{\mathbf{v}}, \underline{\mathbf{w}} \in \mathbb{R}^{(M \times Q)}. \quad (31)$$

Proof. We prove in order the three properties.

- i) **Existence of a solution to \mathcal{L}_Δ^2**
The first property, i.e., the existence of a unique solution to \mathcal{L}_Δ^2 , has already been shown in the proof of its $(M + 1)$ -order accuracy by introducing the operator $\mathcal{J} : \mathbb{R}^{(M \times Q)} \rightarrow \mathbb{R}^{(M \times Q)}$ defined by (16). We showed that for Δt small enough it is a contraction over the space $\mathbb{R}^{(M \times Q)}$ equipped with the infinity norm, so, there exists a unique fixed point of \mathcal{J} , which is the unique solution to \mathcal{L}_Δ^2 .
- ii) **Coercivity-like property of \mathcal{L}_Δ^1**
Let us now consider two generic vectors $\underline{\mathbf{v}}, \underline{\mathbf{w}} \in \mathbb{R}^{(M \times Q)}$

$$\underline{\mathbf{v}} = \begin{pmatrix} \mathbf{v}^1 \\ \vdots \\ \mathbf{v}^m \\ \vdots \\ \mathbf{v}^M \end{pmatrix}, \quad \underline{\mathbf{w}} = \begin{pmatrix} \mathbf{w}^1 \\ \vdots \\ \mathbf{w}^m \\ \vdots \\ \mathbf{w}^M \end{pmatrix}, \quad (32)$$

with \mathbf{v}^m and \mathbf{w}^m for $m = 1, \dots, M$ generic Q -dimensional vectors. From a direct computation, we have

$$\begin{aligned} & \mathcal{L}_\Delta^1(\mathbf{v}) - \mathcal{L}_\Delta^1(\mathbf{w}) \\ &= \begin{pmatrix} \mathbf{v}^1 - \mathbf{u}^0 - \Delta t \beta^1 \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{v}^m - \mathbf{u}^0 - \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{v}^M - \mathbf{u}^0 - \Delta t \beta^M \mathbf{G}(t^0, \mathbf{u}^0) \end{pmatrix} - \begin{pmatrix} \mathbf{w}^1 - \mathbf{u}^0 - \Delta t \beta^1 \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{w}^m - \mathbf{u}^0 - \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{w}^M - \mathbf{u}^0 - \Delta t \beta^M \mathbf{G}(t^0, \mathbf{u}^0) \end{pmatrix} = \begin{pmatrix} \mathbf{v}^1 - \mathbf{w}^1 \\ \vdots \\ \mathbf{v}^m - \mathbf{w}^m \\ \vdots \\ \mathbf{v}^M - \mathbf{w}^M \end{pmatrix}, \end{aligned} \quad (33)$$

i.e., $\mathcal{L}_\Delta^1(\mathbf{v}) - \mathcal{L}_\Delta^1(\mathbf{w}) = \mathbf{v} - \mathbf{w}$. Then,

$$\|\mathcal{L}_\Delta^1(\mathbf{v}) - \mathcal{L}_\Delta^1(\mathbf{w})\|_\infty = \|\mathbf{v} - \mathbf{w}\|_\infty \quad (34)$$

and thus the coercivity-like property of \mathcal{L}_Δ^1 is verified and results in an equality. Again, we remark that \mathbf{u}^0 is given, it is part of the problem and embedded in the operators \mathcal{L}_Δ^1 and \mathcal{L}_Δ^2 .

iii) **Lipschitz-continuity-like condition of $\mathcal{L}_\Delta^1 - \mathcal{L}_\Delta^2$**

Again, we consider a direct computation but focusing, for the sake of compactness, on the Q -dimensional component got for a general m

$$\begin{aligned} & \left[\mathcal{L}_\Delta^{1,m}(\mathbf{v}) - \mathcal{L}_\Delta^{2,m}(\mathbf{v}) \right] - \left[\mathcal{L}_\Delta^{1,m}(\mathbf{w}) - \mathcal{L}_\Delta^{2,m}(\mathbf{w}) \right] \\ &= \mathbf{v}^m - \mathbf{u}^0 - \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) - \mathbf{v}^m + \mathbf{u}^0 + \Delta t \sum_{\ell=0}^M \theta_\ell^m \mathbf{G}(t^\ell, \mathbf{v}^\ell) \\ & \quad - \left[\mathbf{w}^m - \mathbf{u}^0 - \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) - \mathbf{w}^m + \mathbf{u}^0 + \Delta t \sum_{\ell=0}^M \theta_\ell^m \mathbf{G}(t^\ell, \mathbf{w}^\ell) \right] \\ &= \Delta t \sum_{\ell=0}^M \theta_\ell^m (\mathbf{G}(t^\ell, \mathbf{v}^\ell) - \mathbf{G}(t^\ell, \mathbf{w}^\ell)), \end{aligned} \quad (35)$$

where clearly $\mathbf{v}^0 = \mathbf{w}^0 = \mathbf{u}^0$. As we pointed out several times, \mathbf{u}^0 is not an unknown, it is a given vector, it is “part” of the problem and is embedded in the operators. We use \mathbf{v}^0 and \mathbf{w}^0 instead of \mathbf{u}^0 for the sake of compactness. Let us recall that θ_ℓ^m , for $m = 1, \dots, M$ and $\ell = 0, 1, \dots, M$, are fixed constant coefficients independent of Δt , thus bounded in absolute value by a positive constant C_θ , and that $\mathbf{G}(t, \mathbf{u})$ is Lipschitz-continuous with respect to \mathbf{u} uniformly with respect to t with a Lipschitz constant L . By applying the triangular inequality,

we have

$$\begin{aligned}
& \left\| [\mathcal{L}_\Delta^1(\mathbf{v}) - \mathcal{L}_\Delta^2(\mathbf{v})] - [\mathcal{L}_\Delta^1(\mathbf{w}) - \mathcal{L}_\Delta^2(\mathbf{w})] \right\|_\infty \\
&= \Delta t \left\| \sum_{\ell=0}^M \begin{pmatrix} \theta_\ell^1 [\mathbf{G}(t^\ell, \mathbf{v}^\ell) - \mathbf{G}(t^\ell, \mathbf{w}^\ell)] \\ \vdots \\ \theta_\ell^m [\mathbf{G}(t^\ell, \mathbf{v}^\ell) - \mathbf{G}(t^\ell, \mathbf{w}^\ell)] \\ \vdots \\ \theta_\ell^M [\mathbf{G}(t^\ell, \mathbf{v}^\ell) - \mathbf{G}(t^\ell, \mathbf{w}^\ell)] \end{pmatrix} \right\|_\infty \leq \Delta t C_\theta \sum_{\ell=0}^M \left\| \begin{pmatrix} \mathbf{G}(t^\ell, \mathbf{v}^\ell) - \mathbf{G}(t^\ell, \mathbf{w}^\ell) \\ \vdots \\ \mathbf{G}(t^\ell, \mathbf{v}^\ell) - \mathbf{G}(t^\ell, \mathbf{w}^\ell) \\ \vdots \\ \mathbf{G}(t^\ell, \mathbf{v}^\ell) - \mathbf{G}(t^\ell, \mathbf{w}^\ell) \end{pmatrix} \right\|_\infty \\
&= \Delta t C_\theta \sum_{\ell=0}^M \left\| \mathbf{G}(t^\ell, \mathbf{v}^\ell) - \mathbf{G}(t^\ell, \mathbf{w}^\ell) \right\|_{\infty, Q} \leq \Delta t C_\theta \sum_{\ell=0}^M L \left\| \mathbf{v}^\ell - \mathbf{w}^\ell \right\|_{\infty, Q} \leq \Delta t C_\theta L M \left\| \underline{\mathbf{v}} - \underline{\mathbf{w}} \right\|_\infty,
\end{aligned} \tag{36}$$

where the last inequality follows from the fact that $\underline{\mathbf{v}} - \underline{\mathbf{w}}$ contains as components all the vectors $\mathbf{v}^\ell - \mathbf{w}^\ell$ for $\ell = 1, \dots, M$ and from the fact that $\mathbf{v}^0 = \mathbf{w}^0 = \mathbf{u}^0$. This proves the Lipschitz-continuity-like condition of $\mathcal{L}_\Delta^1 - \mathcal{L}_\Delta^2$. For more clarity, we underline that the infinity norm $\|\cdot\|_{\infty, Q}$ is applied to Q -dimensional vectors (and not to $(M \times Q)$ -dimensional vectors like $\|\cdot\|_\infty$). This completes the analysis of the Deferred Correction applied to the context of the systems of ordinary differential equations. \square

2.2 sDeC

The construction of this DeC method makes use of the definition of the subtimenodes introduced for the bDeC method. The main difference is that here we focus on the integration of the system of ODEs in the intervals $[t^{m-1}, t^m]$ rather than $[t^0, t^m]$.

2.2.1 Definition of \mathcal{L}_Δ^2

We start from the exact integration of the system of ODEs in the interval $[t^{m-1}, t^m]$, which would result in

$$\mathbf{u}(t^m) - \mathbf{u}(t^{m-1}) - \int_{t^{m-1}}^{t^m} \mathbf{G}(t, \mathbf{u}(t)) dt = \mathbf{0}, \quad \forall m = 1, \dots, M. \tag{37}$$

Again, in order to get an expression that can actually be used, we replace $\mathbf{G}(t, \mathbf{u}(t))$ with its M -order accurate Lagrange interpolant of degree M associated to the $M + 1$ subtimenodes t^m and replace $\mathbf{u}(t^\ell)$ by \mathbf{u}^ℓ thus getting

$$\mathbf{u}^m - \mathbf{u}^{m-1} - \int_{t^{m-1}}^{t^m} \sum_{\ell=0}^M \mathbf{G}(t^\ell, \mathbf{u}^\ell) \psi^\ell(t) dt = \mathbf{0}, \quad \forall m = 1, \dots, M. \tag{38}$$

Moving the finite sum and the vectors $\mathbf{G}(t^\ell, \mathbf{u}^\ell)$ outside of the integral and performing the exact integration of the Lagrangian polynomial functions $\psi^\ell(t)$ in the subinterval $[t^{m-1}, t^m]$ we get

$$\mathbf{u}^m - \mathbf{u}^{m-1} - \Delta t \sum_{\ell=0}^M \delta_\ell^m \mathbf{G}(t^\ell, \mathbf{u}^\ell) = \mathbf{0}, \quad \forall m = 1, \dots, M, \tag{39}$$

where, just like in the previous case, coefficients δ_ℓ^m are normalized integrals of the Lagrange basis functions independent of Δt .

Our implicit $(M + 1)$ -order accurate operator $\mathcal{L}_\Delta^2 : \mathbb{R}^{(M \times Q)} \rightarrow \mathbb{R}^{(M \times Q)}$ is therefore defined as

$$\mathcal{L}_\Delta^2(\underline{\mathbf{u}}) = \begin{pmatrix} \mathbf{u}^1 - \mathbf{u}^0 - \Delta t \sum_{\ell=0}^M \delta_\ell^1 \mathbf{G}(t^\ell, \mathbf{u}^\ell) \\ \vdots \\ \mathbf{u}^m - \mathbf{u}^{m-1} - \Delta t \sum_{\ell=0}^M \delta_\ell^m \mathbf{G}(t^\ell, \mathbf{u}^\ell) \\ \vdots \\ \mathbf{u}^M - \mathbf{u}^{M-1} - \Delta t \sum_{\ell=0}^M \delta_\ell^M \mathbf{G}(t^\ell, \mathbf{u}^\ell) \end{pmatrix} \text{ with } \underline{\mathbf{u}} = \begin{pmatrix} \mathbf{u}^1 \\ \vdots \\ \mathbf{u}^m \\ \vdots \\ \mathbf{u}^M \end{pmatrix}. \quad (40)$$

2.2.2 Definition of \mathcal{L}_Δ^1

Also in this case the operator \mathcal{L}_Δ^1 is obtained by a first-order approximation in the integration of our initial system of ODEs. Applying the Euler method in the subinterval $[t^{m-1}, t^m]$, we get

$$\mathbf{u}^m - \mathbf{u}^{m-1} - \Delta t \gamma^m \mathbf{G}(t^{m-1}, \mathbf{u}^{m-1}) = \mathbf{0} \quad (41)$$

where $\gamma^m = \frac{t^m - t^{m-1}}{\Delta t}$ are normalized coefficients. The explicit, first-order order operator $\mathcal{L}_\Delta^1 : \mathbb{R}^{(M \times Q)} \rightarrow \mathbb{R}^{(M \times Q)}$ is defined as

$$\mathcal{L}_\Delta^1(\underline{\mathbf{u}}) = \begin{pmatrix} \mathbf{u}^1 - \mathbf{u}^0 - \Delta t \gamma^1 \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{u}^m - \mathbf{u}^{m-1} - \Delta t \gamma^m \mathbf{G}(t^{m-1}, \mathbf{u}^{m-1}) \\ \vdots \\ \mathbf{u}^M - \mathbf{u}^{M-1} - \Delta t \gamma^M \mathbf{G}(t^{M-1}, \mathbf{u}^{M-1}) \end{pmatrix} \text{ with } \underline{\mathbf{u}} = \begin{pmatrix} \mathbf{u}^1 \\ \vdots \\ \mathbf{u}^m \\ \vdots \\ \mathbf{u}^M \end{pmatrix}. \quad (42)$$

2.2.3 sDeC as a perturbation of bDeC

The proofs seen for the previous formulation cannot be extended to this case in a straightforward way, but it is possible to show that the second formulation is actually a perturbation of the first one with no impact on the accuracy. Let us recall here, for more clarity, the updating formulas of the bDeC and of the sDeC methods for the computation of $\mathbf{u}^{m,(p)}$, m -th component of the approximated solution at the iteration p ,

- bDeC

$$\mathbf{u}_b^{m,(p)} = \mathbf{u}^0 + \Delta t \sum_{\ell=0}^M \theta_\ell^m \mathbf{G}(t^\ell, \mathbf{u}_b^{\ell,(p-1)}) \quad (43)$$

- sDeC

$$\mathbf{u}_s^{m,(p)} = \mathbf{u}^0 + \Delta t \sum_{\ell=0}^{m-1} \gamma^{\ell+1} \left(\mathbf{G}(t^\ell, \mathbf{u}_s^{\ell,(p)}) - \mathbf{G}(t^\ell, \mathbf{u}_s^{\ell,(p-1)}) \right) + \Delta t \sum_{\ell=0}^M \theta_\ell^m \mathbf{G}(t^\ell, \mathbf{u}_s^{\ell,(p-1)}). \quad (44)$$

The difference lies in the term

$$\Delta t \sum_{\ell=0}^{m-1} \gamma^\ell \left(\mathbf{G}(t^\ell, \mathbf{u}_s^{\ell,(p)}) - \mathbf{G}(t^\ell, \mathbf{u}_s^{\ell,(p-1)}) \right), \quad (45)$$

which consists in a sum of differences of evaluations of the function \mathbf{G} multiplied by Δt . We will show now why this term can be seen as a perturbation of the updating formula of the first formulation with no impact on the accuracy. This actually depends on the fact that $\mathbf{u}^{\ell,(p)}$ and $\mathbf{u}^{\ell,(p-1)}$ are approximations of the same quantity.

Proposition 2.4 (sDeC accuracy). *The approximation $\mathbf{u}_s^{m,(p)}$ provided by the sDeC (44) is an $O(\Delta t^{p+1})$ perturbation of $\mathbf{u}_b^{m,(p)}$ obtained through the bDeC (43).*

Proof. We will prove it by induction over p and m . The base case of the induction is clearly true as $\mathbf{u}_s^{m,(p)} = \mathbf{u}_b^{m,(p)} = \mathbf{u}^0$ whenever p or m are equal to 0. We focus now on the induction step. We select $p, m \geq 1$ and assume

$$\mathbf{u}_s^{\ell,(k)} = \mathbf{u}_b^{\ell,(k)} + O(\Delta t^{k+1}), \text{ for } \begin{cases} k < p, & \forall \ell = 1, \dots, M, \text{ or} \\ k = p, & \forall \ell \leq m - 1 \end{cases} \quad (46)$$

and we will prove that $\mathbf{u}_s^{m,(p)} = \mathbf{u}_b^{m,(p)} + O(\Delta t^{p+1})$. We start from (44) and, thanks to the induction hypothesis, to the Lipschitz-continuity of \mathbf{G} and by definition of $\mathbf{u}_b^{m,(p)}$ in (43), we have that

$$\begin{aligned} \mathbf{u}_s^{m,(p)} &= \mathbf{u}^0 + \Delta t \sum_{\ell=0}^{m-1} \gamma^{\ell+1} \left(\mathbf{G}(t^\ell, \mathbf{u}_s^{\ell,(p)}) - \mathbf{G}(t^\ell, \mathbf{u}_s^{\ell,(p-1)}) \right) + \Delta t \sum_{\ell=0}^M \theta_\ell^m \mathbf{G}(t^\ell, \mathbf{u}_s^{\ell,(p-1)}) \\ &= \mathbf{u}^0 + \Delta t \sum_{\ell=0}^{m-1} \gamma^{\ell+1} \left(\mathbf{G}(t^\ell, \mathbf{u}_b^{\ell,(p)}) - \mathbf{G}(t^\ell, \mathbf{u}_b^{\ell,(p-1)}) + O(\Delta t^p) \right) \\ &\quad + \Delta t \left(\sum_{\ell=0}^M \theta_\ell^m \mathbf{G}(t^\ell, \mathbf{u}_b^{\ell,(p-1)}) + O(\Delta t^p) \right) \\ &= \mathbf{u}_b^{m,(p)} + \Delta t \sum_{\ell=0}^{m-1} \gamma^{\ell+1} \left(\mathbf{G}(t^\ell, \mathbf{u}_b^{\ell,(p)}) - \mathbf{G}(t^\ell, \mathbf{u}_b^{\ell,(p-1)}) \right) + O(\Delta t^{p+1}). \end{aligned} \quad (47)$$

Thanks again to the Lipschitz-continuity of \mathbf{G} and to the results on the accuracy of the bDeC method, for each $\ell = 1, \dots, m - 1$, we can write

$$\begin{aligned} \left\| \mathbf{G}(t^\ell, \mathbf{u}_b^{\ell,(p)}) - \mathbf{G}(t^\ell, \mathbf{u}_b^{\ell,(p-1)}) \right\|_{\infty, Q} &\leq L \left\| \mathbf{u}_b^{\ell,(p)} - \mathbf{u}_b^{\ell,(p-1)} \right\|_{\infty, Q} \\ &\leq L \left\| \mathbf{u}_\Delta^\ell - \mathbf{u}_\Delta^\ell + O(\Delta t^p) \right\|_{\infty, Q} = O(\Delta t^p), \end{aligned} \quad (48)$$

where \mathbf{u}_Δ^ℓ is the ℓ -th component of \mathbf{u}_Δ , solution to \mathcal{L}_Δ^2 ; further, for $\ell = 0$ we have $\mathbf{G}(t^\ell, \mathbf{u}_b^{\ell,(p)}) - \mathbf{G}(t^\ell, \mathbf{u}_b^{\ell,(p-1)}) = \mathbf{0}$ as the component at the initial subtimenode is always equal to \mathbf{u}^0 . By the previous fact, coming back to (47), we get the thesis

$$\mathbf{u}_s^{m,(p)} = \mathbf{u}_b^{m,(p)} + O(\Delta t^{p+1}). \quad (49)$$

□

3 Continuous Galerkin FEM

Let $\Omega \subset \mathbb{R}^D$ an open regular bounded domain. The general form of a hyperbolic system of balance laws reads

$$\frac{\partial}{\partial t} \mathbf{u}(\mathbf{x}, t) + \operatorname{div}_{\mathbf{x}} \mathbf{F}(\mathbf{u}(\mathbf{x}, t)) = \mathbf{S}(\mathbf{x}, \mathbf{u}(\mathbf{x}, t)), \quad (\mathbf{x}, t) \in \Omega \times \mathbb{R}_0^+, \quad (50)$$

provided with some initial condition $\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x})$ on Ω and some boundary conditions on $\partial\Omega$.

Let us define \mathcal{T}_h a triangulation of $\bar{\Omega}$ and denote with K the general element, which we assume to be convex and closed. Consider the continuous finite element space $V_h = \{g \in C^0(\bar{\Omega}) : g|_K \in \mathbb{P}_M(K) \ \forall K \in \mathcal{T}_h\}$, let $\{\varphi_i\}_{i=1, \dots, I}$ be a basis of V_h such that each φ_i can be associated to a degree of freedom $\mathbf{x}_i \in \bar{\Omega}$ and has support contained in $\mathcal{K}_i := \cup_{K \in K_i} K$, where $K_i := \{K \in \mathcal{T}_h : \mathbf{x}_i \in K\}$. Further, we assume the basis functions normalized in such a way that $\sum_{i=1}^I \varphi_i \equiv 1$. The general form of the semidiscrete formulation of a continuous Galerkin FEM scheme consists in finding a solution $\mathbf{u}_h(\mathbf{x}) = \sum_i \mathbf{c}_i(t) \varphi_i(\mathbf{x})$, with $\mathbf{c}_i(t) \in \mathbb{R}^Q$ at any time t , such that

$$\sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} \left(\int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right) \frac{d}{dt} \mathbf{c}_j(t) + \phi_i(\mathbf{c}(t)) = \mathbf{0}, \quad \forall i = 1, \dots, I, \quad (51)$$

where $\mathbf{S}T_i(\mathbf{u}_h)$ are some stabilization terms and the space residuals $\phi_i(\mathbf{c}(t))$ are defined as

$$\phi_i(\mathbf{c}(t)) = \sum_{K \in K_i} \int_K (\operatorname{div}_{\mathbf{x}} \mathbf{F}(\mathbf{u}_h(\mathbf{x}, t)) - \mathbf{S}(\mathbf{x}, \mathbf{u}_h(\mathbf{x}, t))) \varphi_i(\mathbf{x}) d\mathbf{x} + \mathbf{S}T_i(\mathbf{u}_h), \quad (52)$$

with $\mathbf{c}(t) \in \mathbb{R}^{I \times Q}$ containing as components all the Q -dimensional vectors $\mathbf{c}_i(t)$ associated to the DoFs.

3.1 DeC for CG

In this context, the parameter Δ of the Deferred Correction is the mesh parameter h of the space discretization. We assume CFL conditions on the temporal step size, i.e., $\Delta t \leq Ch$ for some fixed constant $C > 0$. We will implicitly assume the Bernstein polynomials as basis functions; nevertheless, the method can be extended also to other basis functions provided that some constraints concerning the construction of the operator \mathcal{L}_{Δ}^1 , specified in the following, are fulfilled.

3.1.1 Preliminary results

Here, we will present some useful preliminary results that will be used later to prove the first-order accuracy of \mathcal{L}_{Δ}^1 and the Lipschitz-continuity-like condition of $\mathcal{L}_{\Delta}^1 - \mathcal{L}_{\Delta}^2$. In particular, we will prove two propositions, via some intermediate lemmas. We will focus on the Bernstein polynomials; nevertheless the results can be easily extended to other polynomial bases.

Let us consider a general element K , the vector space $\mathbb{P}_M(K)$ of the scalar polynomial functions of degree M defined on it and $u \in \mathbb{P}_M(K)$. We can express u as a linear combination of the Bernstein polynomials $\{\varphi_r\}_{r=1, \dots, R}$ of degree M defined on the element because they are a basis of $\mathbb{P}_M(K)$. We have thus

$$u(\mathbf{x}) = \sum_{r=1}^R c_r \varphi_r(\mathbf{x}), \quad \forall \mathbf{x} \in K, \quad (53)$$

where the scalar coefficients c_r are the Bernstein coefficients associated to the DoFs $\mathbf{x}_r \in K$. Another possibility is to express u in terms of the Lagrange basis functions $\{\hat{\varphi}_r\}_{r=1,\dots,R}$ defined on K which constitute another basis of $\mathbb{P}_M(K)$. Therefore, we can also write

$$u(\mathbf{x}) = \sum_{r=1}^R v_r \hat{\varphi}_r(\mathbf{x}), \quad \forall \mathbf{x} \in K, \quad (54)$$

where the scalar coefficients v_r are the values of u in the DoFs $\mathbf{x}_r \in K$. We define the vector $\mathbf{c} \in \mathbb{R}^R$ of the coefficients of $u \in \mathbb{P}_M(K)$ with respect to the Bernstein basis and the vector $\mathbf{v} \in \mathbb{R}^R$ of the values of u in all the DoFs of K , i.e., the coefficients with respect to the Lagrange basis.

It is always possible to pass from the Bernstein coefficients to the values in the DoFs through the transition matrix T defined as

$$T = \begin{pmatrix} \varphi_1(\mathbf{x}_1) & \varphi_2(\mathbf{x}_1) & \dots & \varphi_R(\mathbf{x}_1) \\ \varphi_1(\mathbf{x}_2) & \varphi_2(\mathbf{x}_2) & \dots & \varphi_R(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(\mathbf{x}_R) & \varphi_2(\mathbf{x}_R) & \dots & \varphi_R(\mathbf{x}_R) \end{pmatrix}. \quad (55)$$

The general element of $T = (T_{ij})_{i,j=1,\dots,R}$ with row index i and column index j is $T_{ij} = \varphi_j(\mathbf{x}_i)$ and we have $\mathbf{v} = T\mathbf{c}$ and $\mathbf{c} = T^{-1}\mathbf{v}$.

Remark 3.1 (Independence of the mesh parameter.). *Neither the matrix T nor its inverse T^{-1} depend on the size of the element K . They just depend on the spatial dimension D and on the degree M . Once we fix D and M , for any specific type of elements, for example the simplices, we have a fixed T and T^{-1} .*

It is clear that the sum of the elements of each row of T is equal to 1, in fact

$$\sum_{j=1}^R T_{ij} = \sum_{j=1}^R \varphi_j(\mathbf{x}_i) = 1, \quad \forall i = 1, \dots, R. \quad (56)$$

This is due to the assumption on the basis functions, which are normalized in such a way that that

$$\sum_{j=1}^R \varphi_j(\mathbf{x}) \equiv 1, \quad \forall \mathbf{x} \in K. \quad (57)$$

Also its inverse T^{-1} enjoys the same property as we will prove in the next lemma.

Lemma 3.1. *The sum of the elements of each row of T^{-1} , inverse of the transition matrix defined in (55), is equal to 1.*

Proof. Let us observe that proving the thesis is equivalent to prove that $T^{-1}\mathbf{1} = \mathbf{1}$ where $\mathbf{1} \in \mathbb{R}^R$ is a vector with all the entries equal to 1. From (56) we have that $T\mathbf{1} = \mathbf{1}$. Thanks to the previous equality, we have that

$$T^{-1}\mathbf{1} = T^{-1}T\mathbf{1} = \mathbf{1} \quad (58)$$

which is the thesis. \square

The previous result will be used to prove the following lemma.

Lemma 3.2. *For any polynomial $u \in \mathbb{P}_M(K)$ such that*

$$u(\mathbf{x}) = \sum_{r=1}^R c_r \varphi_r(\mathbf{x}) = \sum_{r=1}^R v_r \hat{\varphi}_r(\mathbf{x}), \quad \forall \mathbf{x} \in K, \quad (59)$$

where φ_r are the Bernstein polynomials of $\mathbb{P}_M(K)$, c_r the Bernstein coefficients, $\hat{\varphi}_r$ the Lagrange polynomials of $\mathbb{P}_M(K)$ and v_r the Lagrange coefficients, it holds that

$$\sup_{i,j=1,\dots,R} |c_i - c_j| \leq \tilde{C} \sup_{i,j=1,\dots,R} |v_i - v_j|, \quad (60)$$

where $\tilde{C} > 0$ is independent of the size and aspect ratio of K .

Proof. The proof is a straightforward consequence of lemma 3.1. From the fact that $\mathbf{c} = T^{-1}\mathbf{v}$ we know that every Bernstein coefficient c_r can be expressed as a linear combination of the values v_k in the DoFs through the coefficients of the row r of the matrix T^{-1}

$$c_i = \sum_{k=1}^R (T^{-1})_{ik} v_k, \quad c_j = \sum_{k=1}^R (T^{-1})_{jk} v_k \quad (61)$$

and therefore

$$|c_i - c_j| = \left| \sum_{k=1}^R (T^{-1})_{ik} v_k - \sum_{k=1}^R (T^{-1})_{jk} v_k \right|. \quad (62)$$

Now, from lemma 3.1, we know that the coefficients $(T^{-1})_{rk}$ are such that

$$\sum_{k=1}^R (T^{-1})_{rk} = 1 \quad \forall r = 1, \dots, R. \quad (63)$$

This is in particular true for $r = i$ and $r = j$ and so there exist some coefficients $\lambda_{k,\ell}^{i,j}$, depending on i and j , such that (62) can be written as

$$|c_i - c_j| = \left| \sum_{k=1}^R (T^{-1})_{ik} v_k - \sum_{k=1}^R (T^{-1})_{jk} v_k \right| = \left| \sum_{k,\ell=1}^R \lambda_{k,\ell}^{i,j} (v_k - v_\ell) \right|. \quad (64)$$

One simple choice of these coefficients is given by $\lambda_{k,\ell}^{i,j} = \frac{(T^{-1})_{ik} - (T^{-1})_{jk}}{R}$ and a simple computation can be used to prove it. This might lead to suboptimal values of the estimations. The coefficients $\lambda_{k,\ell}^{i,j}$, like the coefficients T_{ij} and $(T^{-1})_{ij}$, do not depend on the size of K , and, thus, they can be bounded by a positive constant C_λ , which depends just on the type of the element considered. Then, thanks to the triangular inequality, (64) gives

$$|c_i - c_j| = \left| \sum_{k,\ell=1}^R \lambda_{k,\ell}^{i,j} (v_k - v_\ell) \right| \leq \sum_{k,\ell=1}^R |\lambda_{k,\ell}^{i,j}| |v_k - v_\ell| \leq C_\lambda \sum_{k,\ell=1}^R |v_k - v_\ell|. \quad (65)$$

Since the number of dimensions D and the degree M are fixed, also R is fixed and so the number of terms in the sum. Therefore, from (65) we get

$$|c_i - c_j| \leq C_\lambda \sum_{k,\ell=1}^R |v_k - v_\ell| \leq \tilde{C} \sup_{i,j=1,\dots,R} |v_i - v_j| \quad (66)$$

for some $\tilde{C} = C_\lambda R^2$ independent of the size of K . \square

This allows to prove the following result.

Lemma 3.3. *For any polynomial $u \in \mathbb{P}_M(K)$ such that $u(\mathbf{x}) = \sum_{r=1}^R c_r \varphi_r(\mathbf{x})$, $\forall \mathbf{x} \in K$, where φ_r are the Bernstein polynomials of $\mathbb{P}_M(K)$ and c_r the Bernstein coefficients, then*

$$\sup_{i,j=1,\dots,R} |c_i - c_j| \leq \tilde{C} h \|\|\nabla_{\mathbf{x}} u\|_1\|_{L^\infty(K)} \quad (67)$$

where \tilde{C} is the positive constant in (60) (and thus independent of the size of K , dependent just on the number of dimensions D , on the degree M and on the type of the element) and h is such that $\text{diam}(K) \leq h$. The norm $\|\cdot\|_1$ is the 1-norm in \mathbb{R}^D , the norm $\|\cdot\|_{L^\infty(K)}$ is the L^∞ norm over K .

Proof. This is a consequence of lemma 3.2, in fact, from basic analysis, we know that for any smooth scalar function $f \in C^1(K)$

$$\sup_{\mathbf{x}, \mathbf{y} \in K} |f(\mathbf{x}) - f(\mathbf{y})| \leq h \|\|\nabla_{\mathbf{x}} f\|_1\|_{L^\infty(K)}, \quad (68)$$

where we remark that K is assumed to be closed. Thus for the polynomial u , thanks to the inequality (60), we have

$$\sup_{i,j=1,\dots,R} |c_i - c_j| \leq \tilde{C} \sup_{i,j=1,\dots,R} |v_i - v_j| \leq \tilde{C} h \|\|\nabla_{\mathbf{x}} u\|_1\|_{L^\infty(K)}, \quad (69)$$

because v_r are the values of u in the DoFs of K . \square

We will continue now with the first proposition of this section, which will be used later in the proofs of the first-order accuracy of \mathcal{L}_Δ^1 and of the Lipschitz-continuity-like condition of $\mathcal{L}_\Delta^1 - \mathcal{L}_\Delta^2$.

Proposition 3.4 (Mass lumping accuracy). *Let us consider a scalar continuous piecewise polynomial function $u \in V_h$. We can write u as a linear combination of the Bernstein polynomials $\{\varphi_i\}_{i=1,\dots,I}$ associated to the tessellation which constitute a basis of V_h , i.e., $u(\mathbf{x}) = \sum_{i=1}^I c_i \varphi_i(\mathbf{x}) \quad \forall \mathbf{x} \in \bar{\Omega}$ with c_i scalar coefficients. Then, we have $\forall i = 1, \dots, I$ that*

$$\left| \sum_{K \in \mathcal{K}_i} c_i \int_K \varphi_i(\mathbf{x}) d\mathbf{x} - \sum_{K \in \mathcal{K}_i} \sum_{\mathbf{x}_j \in K} c_j \int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right| \leq \hat{C} h \|\|\nabla_{\mathbf{x}} u\|_1\|_{L^\infty(\mathcal{K}_i)} \int_{\mathcal{K}_i} |\varphi_i(\mathbf{x})| d\mathbf{x}, \quad (70)$$

with $h = \max_{K \in \mathcal{T}_h} \text{diam}(K)$ and \hat{C} being a constant independent of h , dependent just on the dimension D , on the degree M and on the type of the elements in the mesh.

Proof. We will assume at first all the elements of the tessellation to be of the same type but this hypothesis can be relaxed to the general case with different types of elements.

Let us focus on the left-hand side of (70). Thanks to the normalization (57) of the basis functions and to the fact that the only basis functions that are not identically zero in the element K are the ones associated to the DoFs contained in that element, we can write

$$\left| \sum_{K \in \mathcal{K}_i} c_i \int_K \varphi_i(\mathbf{x}) d\mathbf{x} - \sum_{K \in \mathcal{K}_i} \sum_{\mathbf{x}_j \in K} c_j \int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right| = \left| \sum_{K \in \mathcal{K}_i} \sum_{\mathbf{x}_j \in K} (c_i - c_j) \int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right|. \quad (71)$$

Now, thanks to the triangular inequality, to the fact that the absolute value of the basis functions φ_j can be bounded by a constant C_0 , independent of the size of K , dependent just on the dimension D , on the degree M and on the type of the elements in the tessellation and also to the fact that the number R of DoFs \mathbf{x}_j in each element K is fixed since D and M are fixed, we can write

$$\begin{aligned} & \left| \sum_{K \in \mathcal{K}_i} \sum_{\mathbf{x}_j \in K} (c_i - c_j) \int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right| \leq \sum_{K \in \mathcal{K}_i} \sum_{\mathbf{x}_j \in K} |c_i - c_j| \left| \int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right| \\ & \leq \sum_{K \in \mathcal{K}_i} \sum_{\mathbf{x}_j \in K} \sup_{\mathbf{x}_\ell \in K} |c_i - c_\ell| \int_K |\varphi_i(\mathbf{x})| |\varphi_j(\mathbf{x})| d\mathbf{x} \leq \sum_{K \in \mathcal{K}_i} \sum_{\mathbf{x}_j \in K} C_0 \sup_{\mathbf{x}_\ell \in K} |c_i - c_\ell| \int_K |\varphi_i(\mathbf{x})| d\mathbf{x} \quad (72) \\ & \leq \sum_{K \in \mathcal{K}_i} RC_0 \sup_{\mathbf{x}_\ell \in K} |c_i - c_\ell| \int_K |\varphi_i(\mathbf{x})| d\mathbf{x}. \end{aligned}$$

By applying the previous proposition (67) and from the fact that by definition $\mathcal{K}_i = \cup_{K \in \mathcal{K}_i} K$, we can continue the sequence of inequalities and get

$$\begin{aligned} & \sum_{K \in \mathcal{K}_i} RC_0 \sup_{\mathbf{x}_\ell \in K} |c_i - c_\ell| \int_K |\varphi_i(\mathbf{x})| d\mathbf{x} \leq \sum_{K \in \mathcal{K}_i} RC_0 \tilde{C} h \|\|\nabla_{\mathbf{x}} u\|_1\|_{L^\infty(K)} \int_K |\varphi_i(\mathbf{x})| d\mathbf{x} \\ & \leq RC_0 \tilde{C} h \|\|\nabla_{\mathbf{x}} u\|_1\|_{L^\infty(\mathcal{K}_i)} \sum_{K \in \mathcal{K}_i} \int_K |\varphi_i(\mathbf{x})| d\mathbf{x} = RC_0 \tilde{C} h \|\|\nabla_{\mathbf{x}} u\|_1\|_{L^\infty(\mathcal{K}_i)} \int_{\mathcal{K}_i} |\varphi_i(\mathbf{x})| d\mathbf{x}. \end{aligned} \quad (73)$$

We take $\hat{C} = RC_0 \tilde{C}$ and we have the thesis, in fact, none of R , C_0 and \tilde{C} depend on h , but they just depend on the dimension D , on the degree M and on the type of the elements in the tessellation.

We remark that we assumed that all the elements of the tessellation were of the same type. To deal with the general case in which we have different types of elements we suffice to take \tilde{C} as the maximum of the coefficients \tilde{C} of lemma 3.3 associated to the different types of elements and R as the highest number of degrees of freedom in a single element. \square

Before going ahead let us make some useful observations.

Remark 3.2. *Since the Bernstein basis functions are not negative, we can actually remove the absolute value inside the integral in (70). We left it on purpose to be more general. In fact, it is easy to see that what is proved in this section is actually not limited to the specific case of*

Bernstein polynomials; the results can be easily extended to other polynomial bases, like for example the Lagrange polynomials (for which the matrix T is the identity and the constant $\tilde{C} = 1$) provided that the normalization (57) holds, i.e. $\sum_{i=1}^I \varphi_i(\mathbf{x}) \equiv 1$.

Remark 3.3. The final result (70), which has been proven for a scalar polynomial $u \in V_h$, can be easily extended to the vectorial case by applying it componentwise. If $\mathbf{u} \in V_h^Q$, then we have

$$\mathbf{u}(\mathbf{x}) = \sum_{i=1}^I \mathbf{c}_i \varphi_i(\mathbf{x}) \quad \forall \mathbf{x} \in \bar{\Omega} \quad (74)$$

with $\mathbf{c}_i \in \mathbb{R}^Q \forall i = 1, \dots, I$ being Q -dimensional vectors of coefficients and $\{\varphi_i\}_{i=1, \dots, I}$ the Bernstein basis and it holds that

$$\begin{aligned} & \left\| \sum_{K \in \mathcal{K}_i} \left(\int_K \varphi_i(\mathbf{x}) d\mathbf{x} \right) \mathbf{c}_i - \sum_{K \in \mathcal{K}_i} \sum_{\mathbf{x}_j \in K} \left(\int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right) \mathbf{c}_j \right\|_{\infty} \\ & \leq \hat{C}h \left\| \|\nabla_{\mathbf{x}} \mathbf{u}\|_1 \right\|_{L^\infty(\mathcal{K}_i)} \left\| \int_{\mathcal{K}_i} |\varphi_i(\mathbf{x})| d\mathbf{x} \right\|_{\infty} \quad \forall i = 1, \dots, I \end{aligned} \quad (75)$$

where the norms $\|\cdot\|_1$ and $\|\cdot\|_{L^\infty(\mathcal{K}_i)}$ are applied to each scalar component while the norm $\|\cdot\|_{\infty}$ is on \mathbb{R}^Q .

The key point is that the result (70) is uniform with respect to all the components of \mathbf{u} and so we can easily take the infinity norm of both sides to pass from the scalar to the vectorial case.

We focus now on another intermediate lemma before proving the second and final proposition of this section.

Lemma 3.5. Let $z \in C^1(K)$ and assume that its gradient is bounded in such a way that $\|\|\nabla_{\mathbf{x}} z\|_1\|_{L^\infty(K)} \leq C_g$. Then, for K small enough it holds

$$\|z\|_{L^1(K)} \geq C^* \|z\|_{L^\infty(K)} |K|, \quad (76)$$

with $|K|$ measure of K and C^* a constant dependent on C_g and on $\|z\|_{L^\infty(K)}$ but independent of the size of K .

Proof. As K is closed and $z \in C^1(K)$, then

$$\exists \mathbf{x}^* \in K \text{ s.t. } |z(\mathbf{x}^*)| = \|z\|_{L^\infty(K)} < +\infty. \quad (77)$$

Further, due to the continuity of z , the set B of the points in K for which the absolute value of the function is larger or equal than $\frac{|z(\mathbf{x}^*)|}{2}$ is non-empty and has a strictly positive measure, i.e. $|B| > 0$ with

$$B := \left\{ \mathbf{x} \in K \text{ s.t. } |z(\mathbf{x})| \geq \frac{|z(\mathbf{x}^*)|}{2} \right\}. \quad (78)$$

We try now to find a lower bound for $|B|$ by defining a set $B^* \subseteq B$ whose measure is known; in particular we define

$$B^* := \left\{ \mathbf{x} \in K \text{ s.t. } d(\mathbf{x}, \mathbf{x}^*) \leq \frac{|z(\mathbf{x}^*)|}{2C_g} \right\}. \quad (79)$$

where $d(\cdot, \cdot)$ is the Euclidean distance. Indeed, we have that $B^* \subseteq B$. Let $\tilde{\mathbf{x}} \in B^*$, then by a simple Taylor expansion we get

$$|z(\tilde{\mathbf{x}})| = |z(\mathbf{x}^*) + \nabla_{\mathbf{x}} z(\mathbf{s})(\mathbf{s} - \mathbf{x}^*)| \quad (80)$$

with \mathbf{s} being a point, dependent on $\tilde{\mathbf{x}}$, contained in the segment $S(\tilde{\mathbf{x}}, \mathbf{x}^*)$ connecting $\tilde{\mathbf{x}}$ and \mathbf{x}^* . The triangle inequality gives

$$|z(\tilde{\mathbf{x}})| = |z(\mathbf{x}^*) + \nabla_{\mathbf{x}} z(\mathbf{s})(\mathbf{s} - \mathbf{x}^*)| \geq |z(\mathbf{x}^*)| - |\nabla_{\mathbf{x}} z(\mathbf{s})(\mathbf{s} - \mathbf{x}^*)|. \quad (81)$$

Now, we have that $|\nabla_{\mathbf{x}} z(\mathbf{s})(\mathbf{s} - \mathbf{x}^*)| \leq \frac{|z(\mathbf{x}^*)|}{2}$ because of the regularity assumption on the gradient of z and because $d(\mathbf{s}, \mathbf{x}^*) \leq d(\tilde{\mathbf{x}}, \mathbf{x}^*)$ as \mathbf{s} belongs to the segment $S(\tilde{\mathbf{x}}, \mathbf{x}^*)$. This can be seen by simple computations:

$$|\nabla_{\mathbf{x}} z(\mathbf{s})(\mathbf{s} - \mathbf{x}^*)| \leq \|\nabla_{\mathbf{x}} z\|_{L^\infty(K)} d(\mathbf{s}, \mathbf{x}^*) \leq C_g \frac{|z(\mathbf{x}^*)|}{2C_g} = \frac{|z(\mathbf{x}^*)|}{2}. \quad (82)$$

Coming back to (81) with this information, we can write

$$|z(\tilde{\mathbf{x}})| \geq |z(\mathbf{x}^*)| - |\nabla_{\mathbf{x}} z(\mathbf{s})(\mathbf{s} - \mathbf{x}^*)| \geq |z(\mathbf{x}^*)| - \frac{|z(\mathbf{x}^*)|}{2} = \frac{|z(\mathbf{x}^*)|}{2} \quad (83)$$

and hence $\tilde{\mathbf{x}} \in B$ and $B^* \subseteq B$.

We are able to estimate the measure of B^* providing therefore a lower bound for $|B|$, indeed, by definition, such set is the intersection between K and the ball $B_\rho(\mathbf{x}^*)$ centered in \mathbf{x}^* with radius $\rho := \frac{|z(\mathbf{x}^*)|}{2C_g}$. If the ball $B_\rho(\mathbf{x}^*)$ is entirely contained in K then $B^* = B_\rho(\mathbf{x}^*)$ and its measure is given by $|B^*| = |B_\rho(\mathbf{x}^*)| = C_s \rho^D$ where C_s is the measure of the unitary ball in \mathbb{R}^D . If this does not hold, it is anyway always possible to find a lower bound for the measure of B^* of the type

$$|B^*| \geq \min(C_\alpha \rho^D, |K|) \quad (84)$$

with C_α constant dependent only on the aspect ratio of K but not on its size. Therefore, from the definition of B and from $|B| \geq |B^*| \geq \min(C_\alpha \rho^D, |K|)$, we get

$$\|z\|_{L^1(K)} = \int_K |z(\mathbf{x})| d\mathbf{x} \geq \int_B |z(\mathbf{x})| d\mathbf{x} \geq \frac{|z(\mathbf{x}^*)|}{2} |B| \geq \frac{|z(\mathbf{x}^*)|}{2} |B^*| \geq \frac{|z(\mathbf{x}^*)|}{2} \min(C_\alpha \rho^D, |K|). \quad (85)$$

Now, recalling that $|z(\mathbf{x}^*)| = \|z\|_{L^\infty(K)}$, we have

$$\|z\|_{L^1(K)} \geq \frac{\|z\|_{L^\infty(K)}}{2} |K| \min\left(\frac{C_\alpha \rho^D}{|K|}, 1\right) \quad (86)$$

We define thus $C^* := \frac{1}{2} \min\left(\frac{C_\alpha \rho^D}{|K|}, 1\right)$ and we observe that, since C_α only depends on geometrical properties of K and ρ only depends on z , for K small enough $C^* = \frac{1}{2}$ and we get the thesis. \square

Now, let us generalize this result to the whole domain for piecewise C^1 functions, even discontinuous, by proving the last result of this section.

Proposition 3.6 (Relation between L^∞ and L^1 norms). *Let $z \in \{z \in L^1(\Omega) \text{ s.t. } z|_K \in C^1(K), \forall K \in \mathcal{T}_h\}$ satisfying locally in each element the hypotheses of the previous lemma, i.e. $\|\|\nabla_{\mathbf{x}} z\|_1\|_{L^\infty(K)} \leq C_g$ and K small enough. Assume the mesh to be regular in the sense that for any $i = 1, \dots, I$ it holds that*

$$\int_K |\varphi_i(\mathbf{x})| d\mathbf{x} \leq C_{\mathcal{M}} \int_{\tilde{K}} |\varphi_i(\mathbf{x})| d\mathbf{x}, \quad \forall K, \tilde{K} \in K_i, \quad (87)$$

where $\{\varphi_i\}_{i=1, \dots, I}$ is the basis of V_h given by Bernstein polynomials. Then,

$$\sum_{i=1}^I \|z\|_{L^\infty(\mathcal{K}_i)} \sum_{K \in K_i} \int_K |\varphi_i(\mathbf{x})| d\mathbf{x} \leq \tilde{C}^* \|z\|_{L^1(\Omega)}, \quad (88)$$

where \tilde{C}^* is a positive constant independent of the mesh parameter.

Proof. Let $K^i \in K_i$ be the element such that $\|z\|_{L^\infty(\mathcal{K}_i)} = \|z\|_{L^\infty(K^i)}$; then, using the mesh regularity assumption (87) and the fact that the basis functions are bounded in absolute value by a constant C_0 independent of the mesh parameter, we have

$$\begin{aligned} \sum_{i=1}^I \|z\|_{L^\infty(\mathcal{K}_i)} \sum_{K \in K_i} \int_K |\varphi_i(\mathbf{x})| d\mathbf{x} &= \sum_{i=1}^I \sum_{K \in K_i} \|z\|_{L^\infty(\mathcal{K}_i)} \int_K |\varphi_i(\mathbf{x})| d\mathbf{x} \\ &= \sum_{i=1}^I \sum_{K \in K_i} \|z\|_{L^\infty(K^i)} \int_K |\varphi_i(\mathbf{x})| d\mathbf{x} \\ &\leq \sum_{i=1}^I \sum_{K \in K_i} C_{\mathcal{M}} \|z\|_{L^\infty(K^i)} \int_{K^i} |\varphi_i(\mathbf{x})| d\mathbf{x} \\ &\leq \sum_{i=1}^I \sum_{K \in K_i} C_{\mathcal{M}} C_0 \|z\|_{L^\infty(K^i)} |K^i|. \end{aligned} \quad (89)$$

We apply now the previous lemma 3.5 and, switching the sums over the elements and the DoFs, we get

$$\begin{aligned} \sum_{i=1}^I \sum_{K \in K_i} C_{\mathcal{M}} C_0 \|z\|_{L^\infty(K^i)} |K^i| &\leq \sum_{i=1}^I \sum_{K \in K_i} \frac{C_{\mathcal{M}} C_0}{C^*} \|z\|_{L^1(K^i)} \\ &= \frac{C_{\mathcal{M}} C_0}{C^*} \sum_{K \in \mathcal{T}_h} \sum_{\mathbf{x}_i \in K} \|z\|_{L^1(K^i)} \end{aligned} \quad (90)$$

where C^* is the minimal coefficient of lemma 3.5 among the ones associated to all the elements K^i . If R is the maximal number of DoFs in a single element in the whole mesh, we can continue and write

$$\frac{C_{\mathcal{M}} C_0}{C^*} \sum_{K \in \mathcal{T}_h} \sum_{\mathbf{x}_i \in K} \|z\|_{L^1(K^i)} \leq \frac{R C_{\mathcal{M}} C_0}{C^*} \sum_{K \in \mathcal{T}_h} \sup_{\mathbf{x}_i \in K} \|z\|_{L^1(K^i)}. \quad (91)$$

Now, in (91), each element K in the tessellation is contributing to the sum with the L^1 norm of z over one element K^i among the ones associated to the DoFs $\mathbf{x}_i \in K$. The generic element K^i can

be present in the sum at most a number of times equal to $M_n + 1$ where M_n represents the maximal number of neighbors that an element can have in the tessellation. Hence, we get

$$\frac{RC_{\mathcal{M}}C_0}{C^*} \sum_{K \in \mathcal{T}_n} \sup_{\mathbf{x}_i \in K} \|z\|_{L^1(K^i)} \leq \frac{RC_{\mathcal{M}}C_0}{C^*} (M_n + 1) \|z\|_{L^1(\Omega)}. \quad (92)$$

Observe that none of the coefficients R , $C_{\mathcal{M}}$, C_0 , C^* or M_n depend on the mesh parameter, therefore, by setting $\tilde{C}^* = \frac{RC_{\mathcal{M}}C_0}{C^*} (M_n + 1)$, we get the thesis. \square

Also in this case, we remark that, in the context of Bernstein polynomials, which are non-negative, the absolute value on φ_i is not necessary. We kept it just to be more general. Indeed, all the results can be generalized to other basis functions like the Lagrange polynomials.

3.1.2 Definition of \mathcal{L}_{Δ}^2

The operator \mathcal{L}_{Δ}^2 is the high order implicit operator that we would like to solve. Its definition is not very different from the one seen in the context of the bDeC for ODEs. We introduce the $M + 1$ subtimenodes t^m with $m = 0, \dots, M$ in the interval $[t_n, t_n + \Delta t]$ in which we will consider the approximations of the values of the solution to our system of ODEs. We refer to $\mathbf{c}(t^m)$ as the exact solution in the node t^m and to \mathbf{c}^m as the approximation of the solution in the same node. Clearly, in this case $\mathbf{c}(t^m)$ and \mathbf{c}^m contain as components all the coefficients corresponding to the spatial DoFs, i.e., respectively the vectors $\mathbf{c}_i(t^m)$ of the exact coefficients in the DoFs at the time t^m and the vectors \mathbf{c}_i^m of the approximated ones. As usual, for the first subtimenode we set $\mathbf{c}^0 = \mathbf{c}(t^0) = \mathbf{c}(t_n) = \mathbf{c}_n$ without any approximation. Starting from the exact integration of (51) over $[t^0, t^m]$ and substituting $\phi_i(\mathbf{c}(t))$ with its M -order interpolation in time associated to the $M + 1$ subtimenodes, we get

$$\sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} \left(\int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right) (\mathbf{c}_j^m - \mathbf{c}_j^0) + \Delta t \sum_{\ell=0}^M \theta_{\ell}^m \phi_i(\mathbf{c}^{\ell}) = \mathbf{0}, \quad \forall i = 1, \dots, I \quad \forall m = 1, \dots, M. \quad (93)$$

Therefore, we can define the operator $\mathcal{L}_{\Delta}^2 : \mathbb{R}^{(I \times Q \times M)} \rightarrow \mathbb{R}^{(I \times Q \times M)}$ as

$$\mathcal{L}_{\Delta}^2(\underline{\mathbf{c}}) = (\mathcal{L}_{\Delta,1}^2(\underline{\mathbf{c}}), \mathcal{L}_{\Delta,2}^2(\underline{\mathbf{c}}), \dots, \mathcal{L}_{\Delta,I}^2(\underline{\mathbf{c}})), \quad \forall \underline{\mathbf{c}} \in \mathbb{R}^{(I \times Q \times M)}, \quad (94)$$

where for any i we have

$$\mathcal{L}_{\Delta,i}^2(\underline{\mathbf{c}}) = \begin{pmatrix} \sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} \left(\int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right) (\mathbf{c}_j^1 - \mathbf{c}_j^0) + \Delta t \sum_{\ell=0}^M \theta_{\ell}^1 \phi_i(\mathbf{c}^{\ell}) \\ \vdots \\ \sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} \left(\int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right) (\mathbf{c}_j^m - \mathbf{c}_j^0) + \Delta t \sum_{\ell=0}^M \theta_{\ell}^m \phi_i(\mathbf{c}^{\ell}) \\ \vdots \\ \sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} \left(\int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right) (\mathbf{c}_j^M - \mathbf{c}_j^0) + \Delta t \sum_{\ell=0}^M \theta_{\ell}^M \phi_i(\mathbf{c}^{\ell}) \end{pmatrix}. \quad (95)$$

with the general argument $\underline{\mathbf{c}} \in \mathbb{R}^{(I \times Q \times M)}$ having M components $\mathbf{c}^m \in \mathbb{R}^{(I \times Q)}$ each one associated to a subtimenode and having I components \mathbf{c}_i^m each one associated to a DoF.

The solution $\underline{\mathbf{c}}_{\Delta}$ to $\mathcal{L}_{\Delta}^2(\underline{\mathbf{c}}_{\Delta}) = \mathbf{0}$ is $(M + 1)$ -order accurate in the sense that would contain as components $(M + 1)$ -order accurate approximations of the coefficients which represent the exact solution to (51) in all the subtimenodes t^m $m = 1, \dots, M$. Unfortunately, the problem $\mathcal{L}_{\Delta}^2(\underline{\mathbf{c}}) = \mathbf{0}$ is a huge nonlinear system.

3.1.3 Definition of \mathcal{L}_Δ^1

Performing an Euler approximation in time to numerically solve (51) in $[t^0, t^m]$ we get

$$\sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} \left(\int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right) (\mathbf{c}_j^m - \mathbf{c}_j^0) + \Delta t \beta^m \phi_i(\mathbf{c}^0) = \mathbf{0}, \quad \forall i = 1, \dots, I, \quad \forall m = 1, \dots, M. \quad (96)$$

Further, we perform a first-order mass lumping in space to get a fully explicit approximation formula for \mathbf{c}_i^m

$$C_i (\mathbf{c}_i^m - \mathbf{c}_i^0) + \Delta t \beta^m \phi_i(\mathbf{c}^0) = \mathbf{0}, \quad \forall i = 1, \dots, I \quad \forall m = 1, \dots, M \quad (97)$$

where C_i are constant quantities defined as

$$C_i := \int_\Omega \varphi_i(\mathbf{x}) d\mathbf{x} = \sum_{K \in K_i} \int_K \varphi_i(\mathbf{x}) d\mathbf{x}, \quad \forall i = 1, \dots, I. \quad (98)$$

We assume a choice of the basis functions such that $C_i \neq 0 \forall i$ so that (97) is well-posed. For example, if we choose the Bernstein polynomials, we have $C_i > 0 \forall i$ as the basis functions φ_i are nonnegative. Indeed, \mathbf{c}_i^m got from (97) is a first order approximation of the exact coefficient $\mathbf{c}_i(t^m)$, as proved in the next proposition.

Proposition 3.7 (First order accuracy of (97)). *The solution to (97) is first-order accurate with respect to the exact solution $\mathbf{c}(t)$ to (51) evaluated in all the subtimenodes t^m for $m = 1, \dots, M$.*

Proof. We can equivalently show that if we insert the exact solution to (51) evaluated in all the subtimenodes t^m $m = 1, \dots, M$ into the left-hand side of (97) we get an error $O(\Delta^{D+2})$ where D is the number of spatial dimensions and the parameter Δ is the mesh parameter h of the space discretization. Therefore, we want to prove that

$$C_i (\mathbf{c}_i(t^m) - \mathbf{c}_i^0) + \Delta t \beta^m \phi_i(\mathbf{c}^0) = O(\Delta^{D+2}), \quad \forall i = 1, \dots, I, \quad \forall m = 1, \dots, M. \quad (99)$$

We know that by plugging the exact solution $\mathbf{c}(t)$ in (96) we get an error $O(\Delta^{D+2})$:

$$\sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} \left(\int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right) (\mathbf{c}_j(t^m) - \mathbf{c}_j^0) + \Delta t \beta^m \phi_i(\mathbf{c}^0) = O(\Delta^{D+2}). \quad (100)$$

Hence, instead of (99), we can show that the difference of (99) and (100) is an $O(\Delta^{D+2})$, i.e.,

$$C_i (\mathbf{c}_i(t^m) - \mathbf{c}_i^0) - \sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} \left(\int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right) (\mathbf{c}_j(t^m) - \mathbf{c}_j^0) = O(\Delta^{D+2}). \quad (101)$$

By definition of the coefficients C_i in (98) and the preliminary result (75), we can write

$$\begin{aligned} & \left\| C_i (\mathbf{c}_i(t^m) - \mathbf{c}_i^0) - \sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} \left(\int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right) (\mathbf{c}_j(t^m) - \mathbf{c}_j^0) \right\|_\infty \\ & \leq \hat{C} h \left\| \left\| \nabla_{\mathbf{x}} (\mathbf{u}_h(\mathbf{x}, t^m) - \mathbf{u}_h(\mathbf{x}, t^0)) \right\|_1 \right\|_{L^\infty(\mathcal{K}_i)} \left\| \int_{\mathcal{K}_i} |\varphi_i(\mathbf{x})| d\mathbf{x} \right\|_\infty \end{aligned} \quad (102)$$

where we remark that the internal norms $\|\cdot\|_1$ and $\|\cdot\|_{L^\infty(\mathcal{K}_i)}$ are applied componentwise while the external one, $\|\cdot\|_\infty$, is on \mathbb{R}^Q . From a Taylor expansion it is easy to see that

$$\nabla_{\mathbf{x}}(\mathbf{u}_h(\mathbf{x}, t^m) - \mathbf{u}_h(\mathbf{x}, t^0)) = O(\Delta t). \quad (103)$$

Moreover, $\int_{\mathcal{K}_i} |\varphi_i(\mathbf{x})| d\mathbf{x} = O(\Delta^D)$, hence, we have

$$\left\| C_i (\mathbf{c}_i(t^m) - \mathbf{c}_i^0) - \sum_{K \in \mathcal{K}_i} \sum_{\mathbf{x}_j \in K} \left(\int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right) (\mathbf{c}_j(t^m) - \mathbf{c}_j^0) \right\|_\infty = O(\Delta^{D+2}). \quad (104)$$

□

Directly from (97), we can define the explicit low order operator $\mathcal{L}_\Delta^1 : \mathbb{R}^{(I \times Q \times M)} \rightarrow \mathbb{R}^{(I \times Q \times M)}$ as

$$\mathcal{L}_\Delta^1(\underline{\mathbf{c}}) = (\mathcal{L}_{\Delta,1}^1(\underline{\mathbf{c}}), \mathcal{L}_{\Delta,2}^1(\underline{\mathbf{c}}), \dots, \mathcal{L}_{\Delta,I}^1(\underline{\mathbf{c}})), \quad \forall \underline{\mathbf{c}} \in \mathbb{R}^{(I \times Q \times M)}, \quad (105)$$

where for any i we have

$$\mathcal{L}_{\Delta,i}^1(\underline{\mathbf{c}}) = \begin{pmatrix} C_i (\mathbf{c}_i^1 - \mathbf{c}_i^0) + \Delta t \beta^1 \phi_i(\mathbf{c}^0) \\ \vdots \\ C_i (\mathbf{c}_i^m - \mathbf{c}_i^0) + \Delta t \beta^m \phi_i(\mathbf{c}^0) \\ \vdots \\ C_i (\mathbf{c}_i^M - \mathbf{c}_i^0) + \Delta t \beta^M \phi_i(\mathbf{c}^0) \end{pmatrix}. \quad (106)$$

in which the convention on the indices of the components of the general argument $\underline{\mathbf{c}} \in \mathbb{R}^{(I \times Q \times M)}$ is the same that we had for the operator \mathcal{L}_Δ^2 .

3.1.4 Proof of the properties of \mathcal{L}_Δ^1 and \mathcal{L}_Δ^2

The operators \mathcal{L}_Δ^1 and \mathcal{L}_Δ^2 act from X to Y with $X = Y = \mathbb{R}^{(I \times Q \times M)}$. Let us recall again the hypotheses that are needed in order to apply the Deferred Correction method

i) **Existence of a solution to \mathcal{L}_Δ^2**
 $\exists! \underline{\mathbf{u}}_\Delta \in \mathbb{R}^{(I \times Q \times M)}$ solution of \mathcal{L}_Δ^2 , i.e. such that $\mathcal{L}_\Delta^2(\underline{\mathbf{u}}_\Delta) = \mathbf{0}$;

ii) **Coercivity-like property of \mathcal{L}_Δ^1**
 $\exists \alpha_1 \geq 0$ independent of Δ s.t.

$$\|\mathcal{L}_\Delta^1(\underline{\mathbf{v}}) - \mathcal{L}_\Delta^1(\underline{\mathbf{w}})\|_Y \geq \alpha_1 \|\underline{\mathbf{v}} - \underline{\mathbf{w}}\|_X, \quad \forall \underline{\mathbf{v}}, \underline{\mathbf{w}} \in \mathbb{R}^{(I \times Q \times M)}; \quad (107)$$

iii) **Lipschitz-continuity-like condition of $\mathcal{L}_\Delta^1 - \mathcal{L}_\Delta^2$**
 $\exists \alpha_2 \geq 0$ independent of Δ s.t.

$$\|[\mathcal{L}_\Delta^1(\underline{\mathbf{v}}) - \mathcal{L}_\Delta^2(\underline{\mathbf{v}})] - [\mathcal{L}_\Delta^1(\underline{\mathbf{w}}) - \mathcal{L}_\Delta^2(\underline{\mathbf{w}})]\|_Y \leq \alpha_2 \Delta \|\underline{\mathbf{v}} - \underline{\mathbf{w}}\|_X, \quad \forall \underline{\mathbf{v}}, \underline{\mathbf{w}} \in \mathbb{R}^{(I \times Q \times M)}. \quad (108)$$

We remark that in this context the parameter Δ is the mesh parameter h and that we assume the temporal step size $\Delta t \leq Ch$ for some fixed constant C .

We will not prove the first property, i.e., the existence of a unique solution to \mathcal{L}_Δ^2 , because the proof is identical to the one we had in the ODE case up to the inversion of the mass matrix: from \mathcal{L}_Δ^2 we can define an operator $\mathcal{J} : \mathbb{R}^{(I \times Q \times M)} \rightarrow \mathbb{R}^{(I \times Q \times M)}$ whose fixed points (if any) are solutions to \mathcal{L}_Δ^2 ; further, we can show that for Δ small enough the operator is a contraction over the space $\mathbb{R}^{(I \times Q \times M)}$ equipped with the infinity norm and, hence, there exists a unique fixed point of \mathcal{J} which is the unique solution to \mathcal{L}_Δ^2 .

Before going to the proofs of the other two properties, we need to define the norms adopted on the spaces X and Y . Despite having $X = Y = \mathbb{R}^{(I \times Q \times M)}$ we will equip X and Y with two different norms, differently from what we have done in the ODE case. We will specify the norms after the following useful observation.

Remark 3.4 (Remark on the indices). *The main complication of the proofs is that we have to deal with many indices. We remind that*

- $i = 1, \dots, I$ is referred to the DoFs;
- $q = 1, \dots, Q$ is referred to the components of the approximated solution \mathbf{u}_h to the system of balance laws (50);
- $m = 1, \dots, M$ is referred to the subtimenodes t^m , even if we remark that we also have an initial subtimenode $t^0 = t_n$ in which the quantities are not unknown.

We are already used to the fact that the general element $\underline{\mathbf{c}} \in \mathbb{R}^{(I \times Q \times M)}$ must be thought as a collection of M components $\mathbf{c}^m \in \mathbb{R}^{(I \times Q)}$ $m = 1, \dots, M$. Each component \mathbf{c}^m can be thought as the vector of the coefficients of a vectorial continuous piecewise polynomial function $\mathbf{u}_h(\mathbf{x}, t) = \sum_{i=1}^I \mathbf{c}_i(t) \varphi_i(\mathbf{x})$ evaluated in the subtimenode t^m . In fact, each \mathbf{c}^m is made by I components $\mathbf{c}_i^m \in \mathbb{R}^Q$ with $i = 1, \dots, I$ associated to the DoFs. Finally, each \mathbf{c}_i^m is made by Q components $c_i^{q,m}$ $q = 1, \dots, Q$, scalar coefficients associated to the components of the solution to the system of PDEs that we would like to solve, i.e.,

$$\underline{\mathbf{c}} = \begin{pmatrix} \mathbf{c}^1 \\ \vdots \\ \mathbf{c}^m \\ \vdots \\ \mathbf{c}^M \end{pmatrix} \in \mathbb{R}^{(I \times Q \times M)}, \quad \mathbf{c}^m = \begin{pmatrix} \mathbf{c}_1^m \\ \vdots \\ \mathbf{c}_i^m \\ \vdots \\ \mathbf{c}_I^m \end{pmatrix} \in \mathbb{R}^{(I \times Q)}, \quad \mathbf{c}_i^m = \begin{pmatrix} c_i^{1,m} \\ \vdots \\ c_i^{q,m} \\ \vdots \\ c_i^{Q,m} \end{pmatrix} \in \mathbb{R}^Q. \quad (109)$$

In the proofs, we are going to focus on a single scalar component $q = 1, \dots, Q$ of a single subtimenode $m = 1, \dots, M$ and our results will be uniform with respect to the indices q and m , so we will be able to pass from the scalar results to the desired vectorial results through an infinity norm $\|\cdot\|_\infty$ on $\mathbb{R}^{(Q \times M)}$, similarly to what we did when we passed from (70) to (75) in the preliminary results. Therefore, the norm that we choose for the single component of $\underline{\mathbf{c}} \in X = \mathbb{R}^{(I \times Q \times M)}$ with fixed indices q and m , denoted by $\mathbf{c}^{q,m} \in \mathbb{R}^I$, is the $W_I^{1,1}(\Omega)$ -norm, a discrete version of the classical $W^{1,1}(\Omega)$ -norm. In particular, on a scalar function $u : \Omega \rightarrow \mathbb{R}$ the $W^{1,1}(\Omega)$ -norm is defined as

$$\|u\|_{W^{1,1}(\Omega)} := \|u\|_{L^1(\Omega)} + \sum_{d=1}^D \left\| \frac{\partial}{\partial x_d} u \right\|_{L^1(\Omega)} = \|u\|_{L^1(\Omega)} + \|\nabla_{\mathbf{x}} u\|_{L^1(\Omega)}, \quad (110)$$

from which we define the corresponding discrete norm on \mathbb{R}^I , defined by $\|\cdot\|_{W_I^{1,1}(\Omega)} : \mathbb{R}^I \rightarrow \mathbb{R}_0^+$ as

$$\|\mathbf{c}^{q,m}\|_{W_I^{1,1}(\Omega)} := \left\| \sum_{i=1}^I c_i^{q,m} \varphi_i \right\|_{W^{1,1}(\Omega)}. \quad (111)$$

Using then the classical infinity norm on the space $\mathbb{R}^{Q \times M}$ defined by $\|\cdot\|_{\infty, Q, M} : \mathbb{R}^{Q \times M} \rightarrow \mathbb{R}_0^+$, we introduce the X norm $\|\cdot\|_X : \mathbb{R}^{I \times Q \times M} \rightarrow \mathbb{R}_0^+$ as

$$\|\underline{\mathbf{c}}\|_X := \left\| \left\{ \|\mathbf{c}^{q,m}\|_{W_I^{1,1}(\Omega)} \right\}_{\substack{q=1,\dots,Q \\ m=1,\dots,M}} \right\|_{\infty, Q, M}. \quad (112)$$

Instead, we equip Y with a different norm; we choose for the single component $\tilde{\mathbf{c}}^{q,m} \in \mathbb{R}^I$ with fixed indices q and m of $\tilde{\mathbf{c}} \in Y = \mathbb{R}^{(I \times Q \times M)}$ the 1-norm $\|\cdot\|_{1,I} : \mathbb{R}^I \rightarrow \mathbb{R}_0^+$ defined as

$$\|\tilde{\mathbf{c}}^{q,m}\|_{1,I} := \sum_{i=1}^I |\tilde{c}_i^{q,m}|, \quad (113)$$

then the norm on the whole space $Y = \mathbb{R}^{(I \times Q \times M)}$, $\|\cdot\|_Y : \mathbb{R}^{(I \times Q \times M)} \rightarrow \mathbb{R}_0^+$, is defined by

$$\|\tilde{\mathbf{c}}\|_Y = \left\| \left\{ \|\tilde{\mathbf{c}}^{q,m}\|_{1,I} \right\}_{\substack{q=1,\dots,Q \\ m=1,\dots,M}} \right\|_{\infty, Q, M}. \quad (114)$$

Remark 3.5. *We remark that the initial subtimenode $m = 0$ is not kept into account in the norms (112) and (114) as it is a datum of the problem.*

Remark 3.6 (On the choice of the norms). *The reason of the difference in the norms assumed on X and Y is intuitively due to the following fact. Practically speaking, the elements of X , the arguments of \mathcal{L}_Δ^2 and \mathcal{L}_Δ^1 given respectively by (94) and (105) (and so by (95) and (106)), are the coefficients associated to a vectorial continuous piecewise polynomial function evaluated in the subtimenodes t^m $m = 1, \dots, M$. Therefore, on the space X we take an integral norm for “functions”. Instead, the elements of the space Y , the images of \mathcal{L}_Δ^2 and \mathcal{L}_Δ^1 , are consistent with integrals of the mentioned function associated to the coefficients. In order to guarantee the consistency of the terms in the inequalities to prove and to compare $\|\cdot\|_X$ and $\|\cdot\|_Y$, we must take for Y a norm which does not modify the integral “character” of the components of the elements of the space.*

It is straightforward to prove that (112) and (114) are norms but we will not do it for the sake of brevity. In the context of the proofs of the properties of \mathcal{L}_Δ^2 and \mathcal{L}_Δ^1 , we are going to make use of the two following regularity assumptions.

Assumption 3.8 (Poincaré-like inequality). *We assume that we are working with coefficients regular enough to guarantee that the associated functions \mathbf{g}_h , for some $C_p \geq 0$ independent of Δ , are such that*

$$\|\mathbf{g}_h\|_{W^{1,1}(\Omega)} \leq C_p \|\mathbf{g}_h\|_{L^1(\Omega)}, \quad (115)$$

i.e., we assume that we can control the norm of the gradient of all functions that we will consider with the norm of the functions.

Assumption 3.9 (Smoothness of the space residuals). *We assume the functions ϕ_i defined in (52) to be smooth.*

Finally, the notation in eq. (109) will hold for two generic vectors $\underline{\mathbf{v}}, \underline{\mathbf{w}} \in \mathbb{R}^{(I \times Q \times M)}$ that will be used in the proof.

In order to deal with the single component got for fixed m and q , as we are going to do in a few lines, it is very useful to define here the scalar continuous piecewise polynomial functions

$$v_h^{q,m}(\mathbf{x}) = \sum_{i=1}^I v_i^{q,m} \varphi_i(\mathbf{x}), \quad w_h^{q,m}(\mathbf{x}) = \sum_{i=1}^I w_i^{q,m} \varphi_i(\mathbf{x}) \quad (116)$$

associated to the scalar coefficients $v_i^{q,m}$ and $w_i^{q,m}$ $i = 1, \dots, I$.

Now, we have all the elements that we need in order to handle the proofs of the properties of the two operators.

Proposition 3.10 (Coercivity-like property of \mathcal{L}_Δ^1). *Let $\mathcal{L}_\Delta^1 : X \rightarrow Y$ be the operator defined in (105) and (106), $\underline{\mathbf{v}}, \underline{\mathbf{w}} \in X$ and suppose that assumption 3.8 holds, then $\exists \alpha_1 > 0$ independent of Δ s.t.*

$$\|\mathcal{L}_\Delta^1(\underline{\mathbf{v}}) - \mathcal{L}_\Delta^1(\underline{\mathbf{w}})\|_Y \geq \alpha_1 \|\underline{\mathbf{v}} - \underline{\mathbf{w}}\|_X, \quad \forall \underline{\mathbf{v}}, \underline{\mathbf{w}} \in \mathbb{R}^{(I \times Q \times M)}. \quad (117)$$

Proof. From a direct computation we have, for every $i = 1, \dots, I$, $m = 1, \dots, M$ and $q = 1, \dots, Q$, that

$$\begin{aligned} \mathcal{L}_{\Delta,i}^{1,q,m}(\underline{\mathbf{v}}) - \mathcal{L}_{\Delta,i}^{1,q,m}(\underline{\mathbf{w}}) &= C_i \left(v_i^{q,m} - c_i^{0,q} \right) + \Delta t \beta^m \phi_i^q(\mathbf{c}^0) - C_i \left(w_i^{q,m} - c_i^{0,q} \right) - \Delta t \beta^m \phi_i^q(\mathbf{c}^0) \\ &= C_i (v_i^{q,m} - w_i^{q,m}). \end{aligned} \quad (118)$$

We remark again that \mathbf{c}^0 is known and so also \mathbf{c}_i^0 . We will start by proving the coercivity-like property for a fixed component q and a fixed subtimenode m , i.e., we will prove that the 1-norm of (118) over the indexes $i = 1, \dots, I$ is such that

$$\left\| \mathcal{L}_\Delta^{1,q,m}(\underline{\mathbf{v}}) - \mathcal{L}_\Delta^{1,q,m}(\underline{\mathbf{w}}) \right\|_{1,I} \geq \alpha_1 \|v_h^{q,m} - w_h^{q,m}\|_{W^{1,1}(\Omega)} \quad (119)$$

for some α_1 independent of Δ for all m and q . Recalling the definition (98) of the coefficients $C_i = \int_\Omega \varphi_i(\mathbf{x}) d\mathbf{x}$ and the fact that the Bernstein basis functions are nonnegative, we have

$$\left\| \mathcal{L}_\Delta^{1,q,m}(\underline{\mathbf{v}}) - \mathcal{L}_\Delta^{1,q,m}(\underline{\mathbf{w}}) \right\|_{1,I} = \sum_{i=1}^I |C_i (v_i^{q,m} - w_i^{q,m})| = \sum_{i=1}^I \int_\Omega |(v_i^{q,m} - w_i^{q,m}) \varphi_i(\mathbf{x})| d\mathbf{x}. \quad (120)$$

Using the triangular inequality and recalling the definition (116) of the scalar continuous piecewise polynomial functions $v_h^{q,m}$ and $w_h^{q,m}$, from the previous equation we get

$$\begin{aligned} \left\| \mathcal{L}_\Delta^{1,q,m}(\underline{\mathbf{v}}) - \mathcal{L}_\Delta^{1,q,m}(\underline{\mathbf{w}}) \right\|_{1,I} &\geq \int_\Omega \left| \sum_{i=1}^I (v_i^{q,m} - w_i^{q,m}) \varphi_i(\mathbf{x}) \right| d\mathbf{x} \\ &= \int_\Omega |v_h^{q,m}(\mathbf{x}) - w_h^{q,m}(\mathbf{x})| d\mathbf{x} = \|v_h^{q,m} - w_h^{q,m}\|_{L^1(\Omega)} \\ &\geq \frac{1}{C_p} \|v_h^{q,m} - w_h^{q,m}\|_{W^{1,1}(\Omega)} = \alpha_1 \|v_h^{q,m} - w_h^{q,m}\|_{W^{1,1}(\Omega)}, \end{aligned} \quad (121)$$

where, in the last inequality, we used the Poincaré-like inequality (115) and $\alpha_1 = \frac{1}{C_p}$ with C_p independent of Δ , which is the intermediate result that we wanted to show.

In order to get the final result, it suffices to observe that the previous inequality is uniform with respect to the indices q and m , so, we can take the infinity norm on these indices of both the sides and get

$$\|\mathcal{L}_\Delta^1(\underline{\mathbf{v}}) - \mathcal{L}_\Delta^1(\underline{\mathbf{w}})\|_Y \geq \alpha_1 \|\underline{\mathbf{v}} - \underline{\mathbf{w}}\|_X \quad (122)$$

using the definitions (112) and (114). \square

Proposition 3.11 (Lipschitz-continuity-like condition of $\mathcal{L}_\Delta^1 - \mathcal{L}_\Delta^2$). *Let $\mathcal{L}_\Delta^1, \mathcal{L}_\Delta^2 : X \rightarrow Y$ the operators defined in (105) and (94). Consider $\underline{\mathbf{v}}, \underline{\mathbf{w}} \in X$ regular enough and suppose that assumption 3.9 holds. Then, $\exists \alpha_2 > 0$ independent of Δ s.t.*

$$\|[\mathcal{L}_\Delta^1(\underline{\mathbf{v}}) - \mathcal{L}_\Delta^2(\underline{\mathbf{v}})] - [\mathcal{L}_\Delta^1(\underline{\mathbf{w}}) - \mathcal{L}_\Delta^2(\underline{\mathbf{w}})]\|_Y \leq \alpha_2 \Delta \|\underline{\mathbf{v}} - \underline{\mathbf{w}}\|_X. \quad (123)$$

Proof. Focusing on one DoF $i \in \{1, \dots, I\}$ and on one subtimenode $m \in \{1, \dots, M\}$, we have

$$\begin{aligned} & \left[\mathcal{L}_{\Delta,i}^{1,m}(\underline{\mathbf{v}}) - \mathcal{L}_{\Delta,i}^{2,m}(\underline{\mathbf{v}}) \right] - \left[\mathcal{L}_{\Delta,i}^{1,m}(\underline{\mathbf{w}}) - \mathcal{L}_{\Delta,i}^{2,m}(\underline{\mathbf{w}}) \right] = \\ & C_i (\mathbf{v}_i^m - \mathbf{w}_i^m) - \sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} (\mathbf{v}_j^m - \mathbf{w}_j^m) \int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} - \Delta t \sum_{\ell=0}^M \theta_\ell^m [\phi_i(\mathbf{v}^\ell) - \phi_i(\mathbf{w}^\ell)]. \end{aligned} \quad (124)$$

Just like we did when we proved the coercivity-like property of \mathcal{L}_Δ^1 , we will work on the single component of (124) for fixed $q = 1, \dots, Q$ and $m = 1, \dots, M$, then we will derive the final result on the norms of X and Y by considering the ∞ -norm over the indices q and m .

Let us thus focus on

$$\begin{aligned} & \left[\mathcal{L}_{\Delta,i}^{1,q,m}(\underline{\mathbf{v}}) - \mathcal{L}_{\Delta,i}^{2,q,m}(\underline{\mathbf{v}}) \right] - \left[\mathcal{L}_{\Delta,i}^{1,q,m}(\underline{\mathbf{w}}) - \mathcal{L}_{\Delta,i}^{2,q,m}(\underline{\mathbf{w}}) \right] = C_i (v_i^{q,m} - w_i^{q,m}) \\ & - \sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} (v_j^{q,m} - w_j^{q,m}) \int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} - \Delta t \sum_{\ell=0}^M \theta_\ell^m [\phi_i^q(\mathbf{v}^\ell) - \phi_i^q(\mathbf{w}^\ell)] \end{aligned} \quad (125)$$

where $\phi_i^q(\cdot)$ represents the q -th component of the space residual $\phi_i(\cdot)$. We want to show now that the 1-norm, over all the indices i , of (125), for fixed q and m , is such that

$$\left\| \left[\mathcal{L}_\Delta^{1,q,m}(\underline{\mathbf{v}}) - \mathcal{L}_\Delta^{2,q,m}(\underline{\mathbf{v}}) \right] - \left[\mathcal{L}_\Delta^{1,q,m}(\underline{\mathbf{w}}) - \mathcal{L}_\Delta^{2,q,m}(\underline{\mathbf{w}}) \right] \right\|_{1,I} \leq \alpha_2 \Delta \|\underline{\mathbf{v}} - \underline{\mathbf{w}}\|_X, \quad (126)$$

for some α_2 independent of Δ , from which we will get the final result by taking the infinity norm of the left hand side with respect to the indices q and m . Thanks to the triangular inequality we

have

$$\left\| \left[\mathcal{L}_\Delta^{1,q,m}(\mathbf{v}) - \mathcal{L}_\Delta^{2,q,m}(\mathbf{v}) \right] - \left[\mathcal{L}_\Delta^{1,q,m}(\mathbf{w}) - \mathcal{L}_\Delta^{2,q,m}(\mathbf{w}) \right] \right\|_{1,I} \quad (127)$$

$$= \sum_{i=1}^I \left| \left[\mathcal{L}_{\Delta,i}^{1,q,m}(\mathbf{v}) - \mathcal{L}_{\Delta,i}^{2,q,m}(\mathbf{v}) \right] - \left[\mathcal{L}_{\Delta,i}^{1,q,m}(\mathbf{w}) - \mathcal{L}_{\Delta,i}^{2,q,m}(\mathbf{w}) \right] \right| \quad (128)$$

$$\leq \sum_{i=1}^I \underbrace{\left| \sum_{K \in K_i} (v_i^{q,m} - w_i^{q,m}) \int_K \varphi_i(\mathbf{x}) d\mathbf{x} - \sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} (v_j^{q,m} - w_j^{q,m}) \int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right|}_{=: E_1} \quad (129)$$

$$+ \underbrace{\sum_{i=1}^I \left| \Delta t \sum_{\ell=0}^M \theta_\ell^m [\phi_i^q(\mathbf{v}^\ell) - \phi_i^q(\mathbf{w}^\ell)] \right|}_{=: E_2} \quad (130)$$

recalling the definition of $C_i = \sum_{K \in K_i} \int_K \varphi(\mathbf{x}) d\mathbf{x}$ in (98).

Thanks to the previous inequality, we can deal separately with the two terms of the right hand side, the first one (129) concerning the mass matrix and the second one (130) involving the space residuals, and show that they can be bounded in the following way

$$E_1 \leq C_a \Delta \|\mathbf{v} - \mathbf{w}\|_X, \quad (131)$$

$$E_2 \leq C_b \Delta \|\mathbf{v} - \mathbf{w}\|_X, \quad (132)$$

with C_a and C_b independent of Δ which would give us the desired result.

• **First term concerning the mass matrix**

In order to bound this term, we can directly apply the preliminary result in proposition 3.4 and we get

$$\left| \sum_{K \in K_i} (v_i^{q,m} - w_i^{q,m}) \int_K \varphi_i(\mathbf{x}) d\mathbf{x} - \sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} (v_j^{q,m} - w_j^{q,m}) \int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right| \leq \hat{C} \Delta C_i \|\|\nabla_{\mathbf{x}} (v_h^{q,m} - w_h^{q,m})\|_1\|_{L^\infty(\mathcal{K}_i)}, \quad \forall i = 1, \dots, I, \quad (133)$$

with \hat{C} independent of the mesh parameter $\Delta = h$, dependent just on the number of dimensions D , on the degree M and on the type of the elements in the mesh. From (133) we have

$$E_1 = \sum_{i=1}^I \left| \sum_{K \in K_i} (v_i^{q,m} - w_i^{q,m}) \int_K \varphi_i(\mathbf{x}) d\mathbf{x} - \sum_{K \in K_i} \sum_{\mathbf{x}_j \in K} (v_j^{q,m} - w_j^{q,m}) \int_K \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \right| \leq \hat{C} \Delta \sum_{i=1}^I \|\|\nabla_{\mathbf{x}} (v_h^{q,m} - w_h^{q,m})\|_1\|_{L^\infty(\mathcal{K}_i)} C_i. \quad (134)$$

Thanks to proposition 3.6 taking $z = \|v_h^{q,m} - w_h^{q,m}\|_1$, then (134) can be bounded in the following way

$$\hat{C} \Delta \sum_{i=1}^I \|\|\nabla_{\mathbf{x}} (v_h^{q,m} - w_h^{q,m})\|_1\|_{L^\infty(\mathcal{K}_i)} C_i \leq \hat{C} \Delta \tilde{C}^* \|\|\nabla_{\mathbf{x}} (v_h^{q,m} - w_h^{q,m})\|_1\|_{L^1(\Omega)}. \quad (135)$$

Hence, by definition of the $W^{1,1}(\Omega)$ -norm (110), of the $W_I^{1,1}(\Omega)$ -norm (111) and of the X norm (112), we have

$$\begin{aligned} E_1 &\leq \hat{C} \Delta \tilde{C}^* \left\| \|\nabla_{\mathbf{x}} (v_h^{q,m} - w_h^{q,m})\|_1 \right\|_{L^1(\Omega)} \leq \hat{C} \tilde{C}^* \Delta \|v_h^{q,m} - w_h^{q,m}\|_{W^{1,1}(\Omega)} \\ &\leq C_a \Delta \|v_h^{q,m} - w_h^{q,m}\|_{W^{1,1}(\Omega)} \leq C_a \Delta \|\mathbf{v} - \mathbf{w}\|_X, \end{aligned} \quad (136)$$

with $C_a = \hat{C} \tilde{C}^*$ independent of Δ .

• **Second term involving the space residuals**

By applying the triangular inequality, recalling that θ_ℓ^m are fixed normalized constant coefficients, thus, bounded in absolute value by a positive constant C_θ , and that $\Delta t \leq Ch = C\Delta$ for some fixed constant C , we have

$$E_2 = \sum_{i=1}^I \left| \Delta t \sum_{\ell=0}^M \theta_\ell^m [\phi_i^q(\mathbf{v}^\ell) - \phi_i^q(\mathbf{w}^\ell)] \right| \leq \Delta C C_\theta \sum_{i=1}^I \sum_{\ell=0}^M |\phi_i^q(\mathbf{v}^\ell) - \phi_i^q(\mathbf{w}^\ell)|. \quad (137)$$

From the fact that $\mathbf{v}^0 = \mathbf{w}^0 = \mathbf{c}^0$, we have

$$\Delta C C_\theta \sum_{i=1}^I \sum_{\ell=0}^M |\phi_i^q(\mathbf{v}^\ell) - \phi_i^q(\mathbf{w}^\ell)| \leq \Delta C C_\theta M \sum_{i=1}^I \left\| \{\phi_i^q(\mathbf{v}^m) - \phi_i^q(\mathbf{w}^m)\}_{\substack{q=1,\dots,Q \\ m=1,\dots,M}} \right\|_{\infty, Q, M}. \quad (138)$$

Then, we use the assumption of smoothness of the space residuals $\phi_i(\cdot)$. In particular, we assume the following Lipschitz-continuity-like condition

$$\sum_{i=1}^I \left\| \{\phi_i^q(\mathbf{v}^m) - \phi_i^q(\mathbf{w}^m)\}_{\substack{q=1,\dots,Q \\ m=1,\dots,M}} \right\|_{\infty, Q, M} \leq C_\phi \left\| \|\mathbf{v}_h - \mathbf{w}_h\|_{W^{1,1}(\Omega)} \right\|_{\infty, Q, M} = C_\phi \|\mathbf{v} - \mathbf{w}\|_X \quad (139)$$

with C_ϕ independent of Δ . Using this, from (138) we get

$$E_2 \leq C_b \Delta \|\mathbf{v} - \mathbf{w}\|_X \quad (140)$$

with $C_b = C C_\theta M C_\phi$ independent of Δ , obtaining (132).

Now, that we have proven (131) and (132), the Lipschitz inequality (126) is proven with $\alpha_2 = C_a + C_b$ independent of Δ . Finally, we get the final result by observing that what we have proved holds for any component with fixed indices $q = 1, \dots, Q$ and $m = 1, \dots, M$. So, applying the infinity norm of the left hand side with respect to these indices, we get

$$\begin{aligned} \max_{q,m} \left\| \left[\mathcal{L}_\Delta^{1,q,m}(\mathbf{v}) - \mathcal{L}_\Delta^{2,q,m}(\mathbf{v}) \right] - \left[\mathcal{L}_\Delta^{1,q,m}(\mathbf{w}) - \mathcal{L}_\Delta^{2,q,m}(\mathbf{w}) \right] \right\|_{1,I} \\ = \left\| \left[\mathcal{L}_\Delta^1(\mathbf{v}) - \mathcal{L}_\Delta^2(\mathbf{v}) \right] - \left[\mathcal{L}_\Delta^1(\mathbf{w}) - \mathcal{L}_\Delta^2(\mathbf{w}) \right] \right\|_Y \leq \alpha_2 \Delta \|\mathbf{v} - \mathbf{w}\|_X, \end{aligned} \quad (141)$$

which is the thesis. \square

3.2 Issues with the DeC for CG

We discuss here a negative result seen in the numerical tests even on the monodimensional linear advection equation (LAE) reported in the main document and in many other works, e.g. [2, 6, 7]. The DeC formulation for PDEs with the lumping of the mass matrix does not give the expected formal order of accuracy for space discretizations of order higher than or equal to 4 if one performs the theoretical optimal number of iterations. In this section, we will try to investigate the problem by numerically assessing the impact of the number of iterations P , of the CFL and of the CIP stabilization on higher order derivatives. Before starting, we remark that the loss in the accuracy is not registered in the context of steady problems, indeed, in [1] the expected order of accuracy is obtained with B3 on a nontrivial steady test for the bidimensional Euler equations. Further, one of the authors is involved in a project [4] on some novel CIP stabilizations for the monodimensional SW equations, soon to be submitted, in which the right order of accuracy is obtained for P3, B3 and B4 with the theoretical optimal number of iterations on all the considered steady tests. Therefore, we will focus on the same unsteady test for the monodimensional LAE presented in the main document and, in particular, we will consider P3, B3 and B4 as basis functions and the original formulation of the bDeC for PDEs without interpolations between the iterations as timestepping method. For P3 and B3 we will use, in the context of the CIP stabilization, the same coefficients adopted in the main document, $\delta^{\text{CIP}} = 0.00702$. As the optimal coefficient for B4 is not provided in [5], we will adopt the same coefficient as for B3 and P3. The reference CFL adopted for the tests with B3 and P3 is 0.1, instead, with B4 it is 0.05. Where not specified, such values have been adopted.

3.2.1 Impact of the number of iterations

The numerical results for different number of iterations are reported in fig. 1. In all the cases we can see the same trend: the optimal number of iterations gives order 2, increasing the number of iterations improves the accuracy allowing to reach the formal order. Nevertheless, it is important to notice that many more iterations, with respect to the optimal number, are needed in order to achieve the right order of convergence: 10 for P3, 80 for B3, 320 for B4.

3.2.2 Impact of the CFL

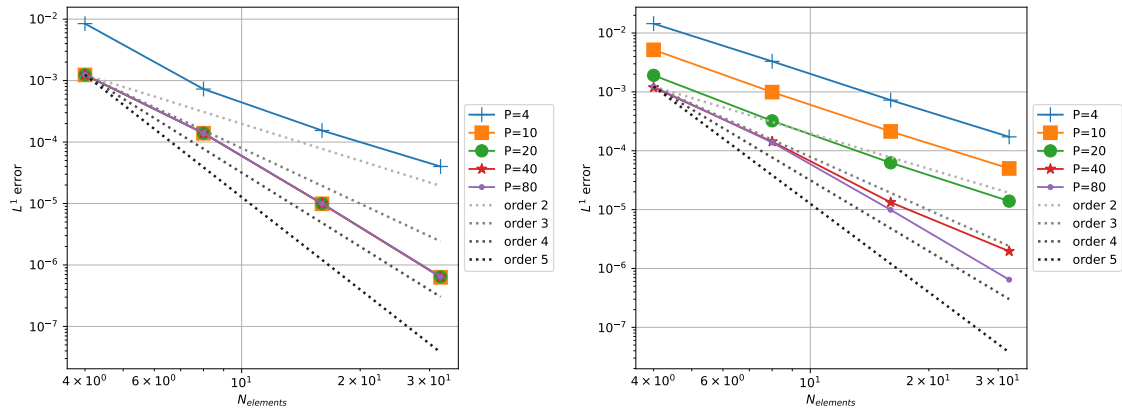
The numerical results for different values of the CFL are reported in fig. 2. Such parameter seems not to have impact on the order. For P3, CFL = 0.1 performs better than CFL = 0.01 and CFL = 0.001; for the other basis functions one gets similar results for the different values of the CFL meaning that spatial error is dominating with respect to the error in time.

3.2.3 Impact of the stabilization on higher order derivatives

The CIP stabilization on the first derivative that we have presented can be actually generalized to keep into account higher order derivatives as in [3]

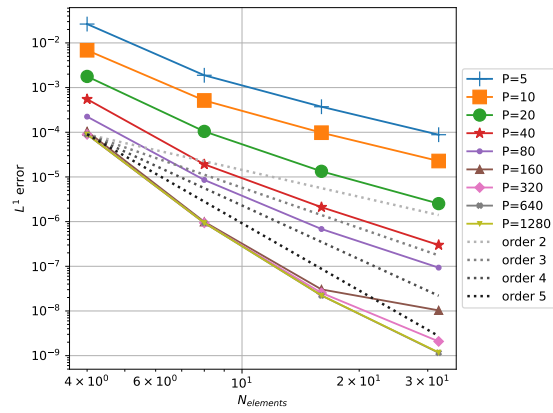
$$\mathbf{ST}_i(\mathbf{u}_h) = \sum_{f \in \mathcal{F}_h} \sum_{r=1}^R \alpha_{f,r}^{\text{CIP}} \int_f \left[\nabla_{\nu_f}^r \varphi_i \right] \cdot \left[\nabla_{\nu_f}^r \mathbf{u}_h \right] d\sigma(\mathbf{x}), \quad \alpha_{f,r}^{\text{CIP}} = \delta_r^{\text{CIP}} \bar{\rho}_f h_f^{2r} \quad (142)$$

where \mathcal{F}_h is the set of the $(D-1)$ -dimensional faces shared by two elements of \mathcal{T}_h , $\nabla_{\nu_f}^r$ is the r -th partial derivative in the direction ν_f normal to the face f and δ_r^{CIP} are constant parameters which



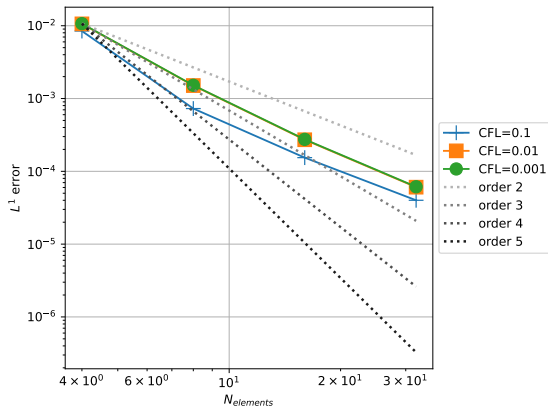
(a) P3

(b) B3

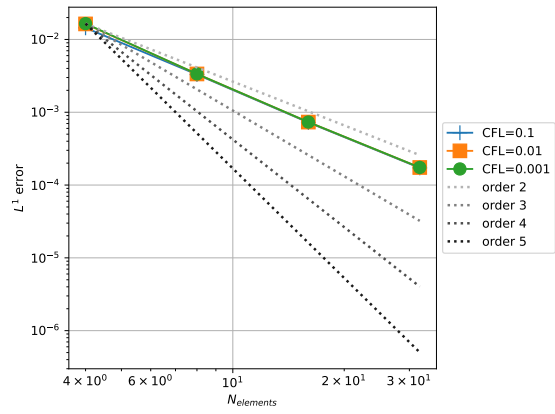


(c) B4

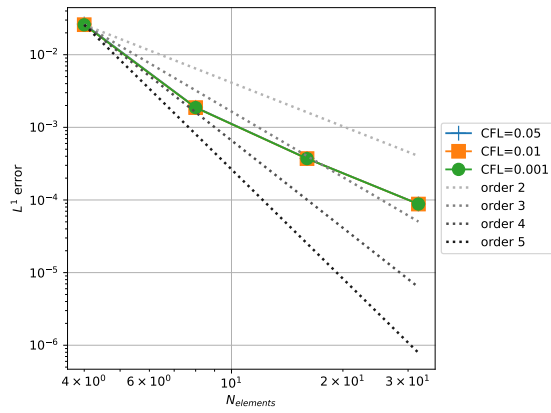
Figure 1: 1D LAE: tests with different numbers of iterations



(a) P3



(b) B3



(c) B4

Figure 2: 1D LAE: tests with different CFLs

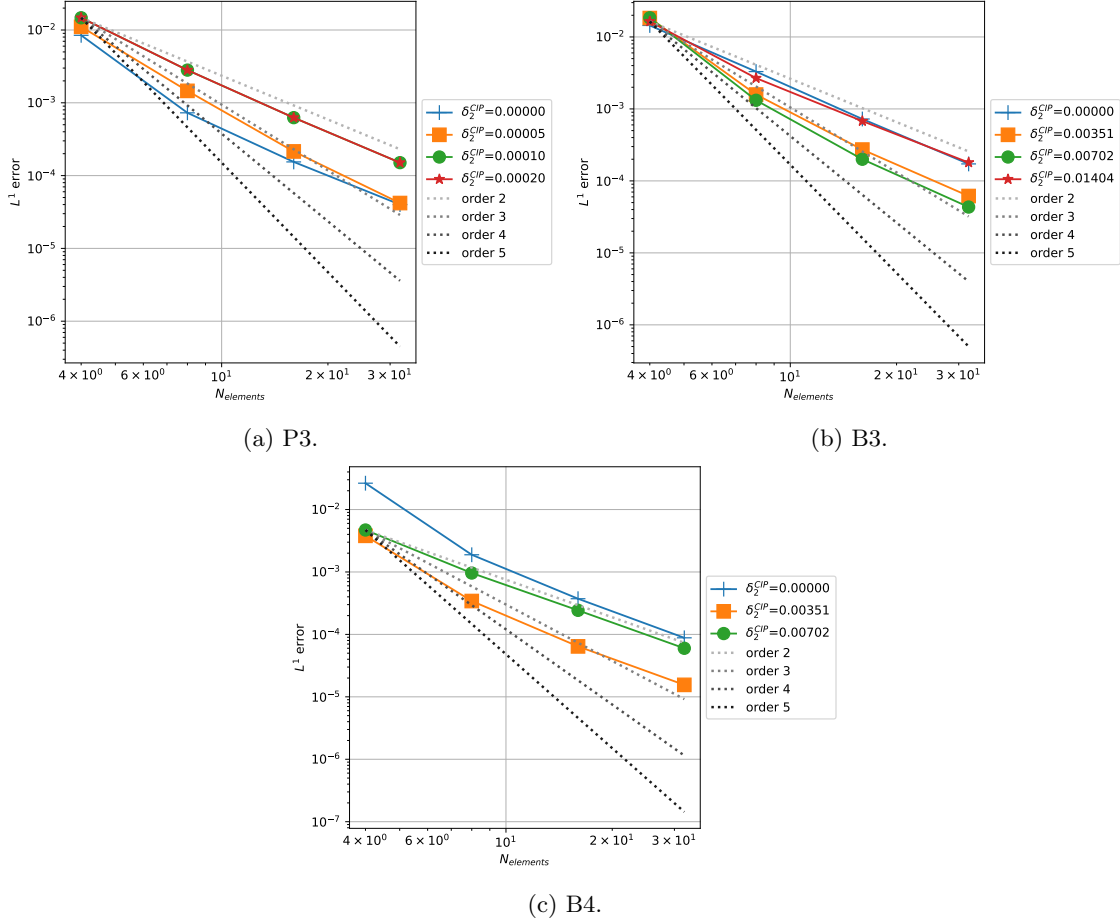


Figure 3: 1D LAE: tests with different stabilization parameters on the second derivative.

must be tuned. We will focus on the stabilization of the first and second derivatives only, $R = 2$. The results obtained with $\delta_1^{\text{CIP}} = 0.00702$ and different values of δ_2^{CIP} are displayed in fig. 3. For B3 and B4, the extra stabilization seems to help in decreasing the errors but still it is not sufficient to achieve the right order of accuracy.

3.2.4 Final remarks

The conclusion of the previous analysis is that, in the context of unsteady problems, with the optimal number of iterations one obtains second order accuracy. Among the three aspects numerically analyzed, only the first one seems to have an effect on the order of accuracy; in particular, many more iterations than expected are needed to reach the formal order. For the moment we do not further investigate this issue, but we have other ideas on how to proceed. First of all, an analysis of the combination of the three parameters studied above could give better results and a linear stability/dispersion analysis, in the style of [5, 6], can help in determining the optimal setting to

achieve the best possible results. Further, higher order derivatives stabilization terms could be taken in consideration hoping for a better stabilization, this has been suggested also in [7, 6]. More in general, other stabilizations other than CIP and OSS could be considered. Moreover, the authors suspect that assumption 3.8 is not verified by the approximations and a stronger estimate on the H^1 norm of the solution of the discrete problem should be provided with weaker hypotheses to guarantee the accuracy results.

We conclude this section remarking that the mentioned problem does not occur with cubature elements also in the DeC framework, which provide accurate and fast results.

4 Vibrating system

Let us consider a general sinusoidal function

$$x(t) = X \cos(\Omega t + \varphi), \quad (143)$$

then we refer to $X \in \mathbb{R}_0^+$ as the amplitude, to $\Omega \in \mathbb{R}^+$ as the frequency and to $\varphi \in [0, 2\pi[$ modulo $2\pi n$ with $n \in \mathbb{Z}$ as the phase.

Let us introduce two general sinusoidal functions

$$x_j(t) = X_j \cos(\Omega t + \varphi_j), \quad \text{for } j = 1, 2, \quad (144)$$

characterized by the same frequency $\Omega > 0$, amplitudes $X_1, X_2 \geq 0$ and phases $\varphi_1, \varphi_2 \in [0, 2\pi[$ modulo $2n\pi$ with $n \in \mathbb{Z}$.

Proposition 4.1. *The sum $x_s(t) = x_1(t) + x_2(t)$ between two sinusoidal functions with the same frequency Ω is another sinusoidal function with the same frequency.*

Proof. If $x_1(t) + x_2(t) = 0$ or at least one between X_1 or X_2 is zero, then the proof is straightforward so let us focus on the case in which $x_1(t) + x_2(t) \neq 0$ and both X_1 and X_2 are different from 0.

From basic trigonometry, we have

$$x_j(t) = X_j \cos(\Omega t + \varphi_j) = X_j [\cos(\Omega t) \cos(\varphi_j) - \sin(\Omega t) \sin(\varphi_j)], \quad \text{for } j = 1, 2, \quad (145)$$

then

$$x_s(t) = x_1(t) + x_2(t) = A \cos(\Omega t) - B \sin(\Omega t), \quad (146)$$

$$\text{with } A := X_1 \cos(\varphi_1) + X_2 \cos(\varphi_2), \quad B := X_1 \sin(\varphi_1) + X_2 \sin(\varphi_2). \quad (147)$$

We consider now the point $(A, B) \in \mathbb{R}^2$, different from $(0, 0)$ by assumption, and the induced vector of length $X_s = \sqrt{A^2 + B^2}$ and phase $\varphi_s = \angle(A, B)$, so that $A = X_s \cos(\varphi_s)$ and $B = X_s \sin(\varphi_s)$. By definition of such vector, (146) can be recast as

$$x_1(t) + x_2(t) = X_s \cos(\varphi_s) \cos(\Omega t) - X_s \sin(\varphi_s) \sin(\Omega t) = X_s \cos(\Omega t + \varphi_s), \quad (148)$$

which completes the proof. \square

We introduce now a bijection \mathcal{S} from the quotient set of the sinusoidal functions with a fixed frequency Ω defined by (X, φ) , in which we identify all the functions characterized by $X = 0$, onto the complex plane

$$\mathcal{S}(x(t)) = \mathcal{S}(X, \varphi) = \begin{cases} X e^{i\varphi} & \text{if } X \neq 0 \\ 0 & \text{if } X = 0 \end{cases}. \quad (149)$$

The complex number $\overline{X} := \mathcal{S}(x(t))$ is called phasor associated to the sinusoidal function $x(t)$.

Proposition 4.2. *If we have two sinusoidal functions $x_1(t), x_2(t)$ with the same frequency Ω then the phasor \overline{X}_s associated to the sum $x_s(t)$ of the two sinusoidal functions is the sum of the phasors $\overline{X}_1, \overline{X}_2$ associated to the single sinusoidal functions.*

Proof. The phasors related to the sinusoidal functions (144) are given by

$$\overline{X}_j = X_j e^{i\varphi_j} = X_j [\cos(\varphi_j) + i \sin(\varphi_j)], \text{ for } j = 1, 2. \quad (150)$$

If one between X_1 or X_2 is zero then the proof is straightforward therefore we focus on the case in which they are both different from 0. Further, we assume for the moment that $x_1(t) + x_2(t) \neq 0$. The sum of the phasors gives

$$\begin{aligned} \overline{X}_r &= \overline{X}_1 + \overline{X}_2 \\ &= [X_1 \cos(\varphi_1) + X_2 \cos(\varphi_2)] + i [X_1 \sin(\varphi_1) + X_2 \sin(\varphi_2)] = A + iB \end{aligned} \quad (151)$$

with A and B defined exactly as in (147) leading to

$$\overline{X}_r = X_r e^{i\varphi_r} \quad (152)$$

with $X_r = X_s$ and $\varphi_r = \varphi_s$ with X_s and φ_s defined from the phasor associated to $x_s(t)$.

If $x_1(t) + x_2(t) = 0$, by simple considerations, we must have $X_2 = X_1$ and $\varphi_2 = \varphi_1 + \pi$ modulo 2π , which leads to

$$\overline{X}_1 = X_1 e^{i\varphi_1}, \quad \overline{X}_2 = X_1 e^{i(\varphi_1 + \pi)} = -\overline{X}_1. \quad (153)$$

Then, we clearly have $\overline{X}_1 + \overline{X}_2 = 0$. Indeed, also the phasor \overline{X}_s associated to the sum is 0 and this completes the proof. \square

It is clear that if we have a sinusoidal function $x(t) = X \cos(\Omega t + \varphi)$ then its derivative in time is still a sinusoidal function with the same frequency

$$x'(t) = -\Omega X \sin(\Omega t + \varphi) = \Omega X \cos\left(\Omega t + \varphi + \frac{\pi}{2}\right). \quad (154)$$

Then the phasor \overline{X}' associated to the derivative in time $x'(t)$ is

$$\overline{X}' = \Omega X e^{i(\varphi + \frac{\pi}{2})} = i\Omega X e^{i\varphi} = i\Omega \overline{X}. \quad (155)$$

By the same argument we have that the phasor \overline{X}'' associated to the second derivative in time $x''(t)$ is

$$\overline{X}'' = i\Omega \overline{X}' = i\Omega(i\Omega \overline{X}) = -\Omega^2 \overline{X}. \quad (156)$$

We consider the scalar ODE

$$\begin{cases} my'' + ry' + ky = F \cos(\Omega t + \varphi), & t \in \mathbb{R}_0^+ \\ y(0) = A, \\ y'(0) = B, \end{cases} \quad (157)$$

with the real nonnegative constants $m, k, \Omega > 0$ and $r, F \geq 0$ with $\varphi \in [0, 2\pi[$ modulo $2\pi n$ with $n \in \mathbb{Z}$. The solution to (157) is given by

$$y(t) = y_h(t) + y_p(t) \quad (158)$$

where $y_h(t)$ is a solution to the homogeneous equation and $y_p(t)$ is a solution to the whole equation.

We first focus on the homogeneous problem

$$my'' + ry' + ky = 0 \quad (159)$$

and we look for a solution in the form $y(t) = Ae^{\lambda t}$ which is nontrivial and so we assume $A \neq 0$. We substitute it in the homogeneous equation and we get

$$(m\lambda^2 + r\lambda + k) Ae^{\lambda t} = 0 \quad (160)$$

and since $Ae^{\lambda t} \neq 0 \forall t \in \mathbb{R}_0^+$ because $A \neq 0$ then we get the characteristic equation

$$\lambda^2 + \alpha\lambda + \beta = 0 \quad (161)$$

with $\alpha = \frac{r}{m} \geq 0$ and $\beta = \frac{k}{m} > 0$. The roots are given by

$$\lambda_{1,2} = \frac{1}{2} \left(-\alpha \pm \sqrt{\alpha^2 - 4\beta} \right) \quad (162)$$

and, depending on the parameters of the problem, we have three possibilities

1. $\lambda_1 \neq \lambda_2$, real, negative and different if $\alpha > 2\sqrt{\beta} \Leftrightarrow r > 2\sqrt{km}$;
2. $\lambda_1 = \lambda_2 = \lambda$, real, negative and coincident if $\alpha = 2\sqrt{\beta} \Leftrightarrow r = 2\sqrt{km}$;
3. $\lambda_{1,2} = \alpha \pm i\omega$, complex and conjugate with negative real part if $\alpha < 2\sqrt{\beta} \Leftrightarrow r < 2\sqrt{km}$.

Thus, the solution to our homogeneous ODE is

$$y_h(t) = \begin{cases} C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}, & \text{if } \alpha > 2\sqrt{\beta} \Leftrightarrow r > 2\sqrt{km}, \\ C_1 e^{\lambda t} + C_2 t e^{\lambda t}, & \text{if } \alpha = 2\sqrt{\beta} \Leftrightarrow r = 2\sqrt{km}, \\ e^{-\frac{\alpha}{2}t} (C_1 \cos(\omega t) + C_2 \sin(\omega t)), & \text{if } \alpha < 2\sqrt{\beta} \Leftrightarrow r < 2\sqrt{km}. \end{cases} \quad (163)$$

Now, we focus on the whole ODE (157) and we assume a sinusoidal solution of the type $y_p = Y_p \cos(\Omega t + \psi)$, we substitute it in (157) and we solve the equation in the space of the phasors. Recalling the expression of the phasors associated to the first and the second derivatives of a sinusoidal function given by (155) and (156) we have

$$-m\Omega^2 \bar{Y}_p + i\Omega r \bar{Y}_p + k \bar{Y}_p = F e^{i\varphi}. \quad (164)$$

Then

$$\bar{Y}_p = \frac{F e^{i\varphi}}{-m\Omega^2 + k + i\Omega r}, \quad (165)$$

from which we get

$$Y_p = \frac{F}{\sqrt{(-m\Omega^2 + k)^2 + \Omega^2 r^2}}, \quad \psi = \varphi - \arg(-m\Omega^2 + k + i\Omega r), \quad (166)$$

where by $\arg(\cdot)$ we denote the phase of the argument up to $2n\pi$ with $n \in \mathbb{Z}$. Once we compute \overline{Y}_p , we automatically get the unique associated sinusoidal function $y_p(t) = Y_p \cos(\Omega t + \psi)$.

So, the final solution to our ODE (157) is $y(t) = y_h(t) + y_p(t)$, where $y_h(t)$ is given by (163) and $y_p(t)$ is a sinusoidal function whose amplitude and phase are given by (166).

The two constants C_1 and C_2 in $y_h(t)$ are computed by imposing the initial conditions $y(0) = A$ and $y'(0) = B$ and solving the resulting 2 by 2 linear system.

References

- [1] Rémi Abgrall and Davide Torlo. High order asymptotic preserving deferred correction implicit-explicit schemes for kinetic models. *SIAM Journal on Scientific Computing*, 42(3):B816–B845, 2020.
- [2] Rémi Abgrall. High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices. *J. Sci. Comput.*, 73(2-3):461–494, 2017.
- [3] Mats G Larson and Sara Zahedi. Stabilization of high order cut finite element methods on surfaces. *IMA Journal of Numerical Analysis*, 40(3):1702–1745, 2020.
- [4] Lorenzo Micalizzi, Mario Ricchiuto, and Rémi Abgrall. Novel well-balanced arbitrary high order continuous interior penalty stabilization techniques for continuous galerkin fem and residual distribution. *in preparation*, 2022.
- [5] Sixtine Michel, Davide Torlo, Mario Ricchiuto, and Rémi Abgrall. Spectral analysis of continuous FEM for hyperbolic PDEs: influence of approximation, stabilization, and time-stepping. *Journal of Scientific Computing*, 89(2):1–41, 2021.
- [6] Sixtine Michel, Davide Torlo, Mario Ricchiuto, and Rémi Abgrall. Spectral analysis of high order continuous FEM for hyperbolic PDEs on triangular meshes: influence of approximation, stabilization, and time-stepping. *arXiv preprint arXiv:2206.06150*, 2022.
- [7] Abgrall Rémi, Bacigaluppi Paola, and Tokareva Svetlana. High-order residual distribution scheme for the time-dependent euler equations of fluid dynamics. *Computers & Mathematics with Applications*, 78(2):274–297, 2019. Proceedings of the Eight International Conference on Numerical Methods for Multi-Material Fluid Flows (MULTIMAT 2017).