Universität Zürich

Institut für Mathematik

Lorenzo Micalizzi

Deferred Correction: from simple integration of systems of ordinary differential equations to explicit time-integration in Residual Distribution and polynomial Continuous and Discontinuous Galerkin FEM schemes for hyperbolic systems of balance laws

Prof. Remi Abgrall

November 2020

Contents

List of Figures

Chapter 1

Abstract

This work is an attempt to give a systematic introduction to the Deferred Correction method[1], or simply DeC, and its application in the context of the numerical methods for hyperboolic systems of balance laws. The Deferred Correction approach provides an iterative procedure to get arbitrary high order accurate integration schemes for systems of ordinary differential equations. The original method has been introduced by Alouk Dutt, Leslie Greengard and Vladimir Rokhlin in their work "Spectral Deferred Correction Methods for Ordinary Differential Equations" published in "BIT Numerical Mathematics", volume 40, pages 241-266 in 2000. It can also be applied in the context of Residual Distribution and polynomial Continuous and Discontinuous Galerkin Finite Elements methods for hyperbolic systems of balance laws to get an explicit formulation and avoid the inversion of mass matrices. The application of the Deferred Correction method to the Residual Distribution framework is due to Remi Abgrall in "High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices" published in the "Journal of Scientific Computing", volume 73, pages 461–494 in 2017. The structure of this script is the following:

- We start by introducing the Deferred Correction as a general method to approximate the solution of a general "tough" operator $\mathcal{L}^2_\Delta$ with an arbitrary high order precision through an iterative procedure which makes use of the operator $\mathcal{L}^1_\Delta$ which is a "naive" operator. In the applications the "tough" operator $\mathcal{L}^2_\Delta$ is an implicit high-order operator difficult to solve while the "naive" operator $\mathcal{L}^1_\Delta$ is an explicit low-order operator easy to solve.

- Then we apply this method to the context of the integration of systems of ordinary differential equations.

- In order to apply the Deferred Correction to some numerical methods for solving hyperbolic systems of balance laws, we present a small introduction of the analytical problem. We derive the balance laws and we introduce their weak formulation in order to overcome the limits of the classical one.

[1]The method is sometimes referred as Defect Correction.

- Thus we introduce from a general point of view the semidiscrete formulation which is a framework common to many numerical methods for hyperbolic systems of balance laws (like Residual Distribution and Galerkin Finite Elements) and focus on the problem of the inversion of the mass matrix.
- Hence we present the Residual Distribution and the stabilized Continuous Galerkin Finite Elements methods and we show the connection between these approaches in the case of polynomial approximations. Thus we see how to apply the Deferred Correction in the context of these methods for which avoiding the inversion of the mass matrix without spoiling the accuracy of the scheme plays a crucial role.
- Therefore we introduce the Discontinuous Galerkin Finite Elements approach and show how the Deferred Correction can be enforced also in this case again considering polynomial approximations.
- In the end we show how the Deferred Correction method, as an ODE solver, can be seen as a particular Runge-Kutta scheme.

Chapter 2

The Deferred Correction as a general procedure in an abstract framework

We will first introduce the Deferred Correction in an abstract context. Assume that we have two general operators depending on a parameter[1] Δ between two normed vector spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$

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

Observation 1. *Even if we are still in an abstract context and not in the specific case of an evolution problem (like for example an ordinary differential equation) it is useful in order to make things clearer to give an idea of the objects we are working with. Imagine that we want to solve numerically a Cauchy problem for a system of ODEs, then $\mathcal{L}^2_\Delta$ is a high-order implicit operator and $\mathcal{L}^1_\Delta$ is an explicit low-order operator (for example the operator that we get by using the Euler approximation). We would like to solve $\mathcal{L}^2_\Delta$ i.e. finding $\mathbf{u} \in X$ such that $\mathcal{L}^2_\Delta(\underline{\mathbf{u}}) = \mathbf{0}_Y$ but this is not so easy given the implicit nature of the operator. On the other hand the other exlicit operator $\mathcal{L}^1_\Delta$ is very easy to solve[2] but it is a low-order operator.*

This observation was meant to provide a better understanding of the meaning of the operators $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$ and of the direction that we are going to take. Let's now go back to our abstract framework, we will focus more in detail later on the application to the evolution problems. For the moment just imagine that we have to solve $\mathcal{L}^2_\Delta$ (because it

[1] This parameter represents the time-step in the context of the integration of ordinary differential equations, while it represents the characteristic mesh size in the application to hyperbolic systems of balance laws. The esteems that we will get will be strictly depending on this parameter which has to be fixed in advance. In other words we are able to prove the arbitrary high order consistency for a fixed Δ and what happens in the mesh refinment is still unclear from a theoretical point of view even if numerical experiments seem to confirm the desired order of convergence.

[2] More in general it is easy to solve $\mathcal{L}^1_\Delta = \underline{\mathbf{v}}$ with $\underline{\mathbf{v}} \in Y$ i.e. finding $\underline{\mathbf{u}} \in X$ such that $\mathcal{L}^1_\Delta(\underline{\mathbf{u}}) = \underline{\mathbf{v}}$.

provides a more accurate solution to a more general problem that we would like to solve) but this operator is difficult to solve. We would like to solve $\mathcal{L}^1_\Delta$ instead of $\mathcal{L}^2_\Delta$ because it is easier to solve (but its solution is not enough accurate with respect to the one our general problem). In the next theorem we will provide a receipt to get an arbitrary high order approximation of the solution of $\mathcal{L}^2_\Delta$ in an explicit way by combining the operators $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$ through an iterative procedure.

Theorem 1. ***Deferred Correction***
If the following hypoteses hold

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

ii) ***Sub-coercivity of*** $\mathcal{L}^1_\Delta$
$\exists \alpha_1 \geq 0$ *independent of* Δ *s.t.*

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

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

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

then if we consider

$$\mathcal{L}^1_\Delta(\underline{\mathbf{u}}^{(p)}) = \mathcal{L}^1_\Delta(\underline{\mathbf{u}}^{(p-1)}) - \mathcal{L}^2_\Delta(\underline{\mathbf{u}}^{(p-1)}) \quad p = 1, 2, \ldots, P \tag{2.1}$$

we have that

$$\left\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}_\Delta\right\|_X \leq \left(\Delta \frac{\alpha_2}{\alpha_1}\right)^P \left\|\underline{\mathbf{u}}^{(0)} - \underline{\mathbf{u}}_\Delta\right\|_X . \tag{2.2}$$

Proof. By using the sub-coercivity of $\mathcal{L}^1_\Delta$ we have

$$\left\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}_\Delta\right\|_X \leq \frac{1}{\alpha_1} \left\|\mathcal{L}^1_\Delta(\underline{\mathbf{u}}^{(P)}) - \mathcal{L}^1_\Delta(\underline{\mathbf{u}}_\Delta)\right\|_Y . \tag{2.3}$$

Let's focus on the right hand side of this inequality. By definition of $\mathcal{L}^1_\Delta(\underline{\mathbf{u}}^{(p)})$ for $p = 1, 2, \ldots, P$ in equation 2.1, we have

$$\frac{1}{\alpha_1} \left\|\mathcal{L}^1_\Delta(\underline{\mathbf{u}}^{(P)}) - \mathcal{L}^1_\Delta(\underline{\mathbf{u}}_\Delta)\right\|_Y = \frac{1}{\alpha_1} \left\|\mathcal{L}^1_\Delta(\underline{\mathbf{u}}^{(P-1)}) - \mathcal{L}^2_\Delta(\underline{\mathbf{u}}^{(P-1)}) - \mathcal{L}^1_\Delta(\underline{\mathbf{u}}_\Delta)\right\|_Y$$

and since $\underline{\mathbf{u}}_\Delta$ is the solution of $\mathcal{L}^2_\Delta$ we have that $\mathcal{L}^2_\Delta(\underline{\mathbf{u}}_\Delta) = \mathbf{0}_Y$ and we can add it on the right hand side and get

$$\frac{1}{\alpha_1} \left\|\mathcal{L}^1_\Delta(\underline{\mathbf{u}}^{(P-1)}) - \mathcal{L}^2_\Delta(\underline{\mathbf{u}}^{(P-1)}) - \mathcal{L}^1_\Delta(\underline{\mathbf{u}}_\Delta)\right\|_Y = \frac{1}{\alpha_1} \left\|\mathcal{L}^1_\Delta(\underline{\mathbf{u}}^{(P-1)}) - \mathcal{L}^2_\Delta(\underline{\mathbf{u}}^{(P-1)}) - \mathcal{L}^1_\Delta(\underline{\mathbf{u}}_\Delta) + \mathcal{L}^2_\Delta(\underline{\mathbf{u}}_\Delta)\right\|_Y .$$

Thus the initial inequality 2.3 becomes

$$\left\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}_\Delta\right\|_X \leq \frac{1}{\alpha_1}\left\|\left[\mathcal{L}^1_\Delta(\underline{\mathbf{u}}^{(P-1)}) - \mathcal{L}^2_\Delta(\underline{\mathbf{u}}^{(P-1)})\right] - \left[\mathcal{L}^1_\Delta(\underline{\mathbf{u}}_\Delta) - \mathcal{L}^2_\Delta(\underline{\mathbf{u}}_\Delta)\right]\right\|_Y. \quad (2.4)$$

Thanks to the Lipschitz-continuity-like condition we can write

$$\frac{1}{\alpha_1}\left\|\left[\mathcal{L}^1_\Delta(\underline{\mathbf{u}}^{(P-1)}) - \mathcal{L}^2_\Delta(\underline{\mathbf{u}}^{(P-1)})\right] - \left[\mathcal{L}^1_\Delta(\underline{\mathbf{u}}_\Delta) - \mathcal{L}^2_\Delta(\underline{\mathbf{u}}_\Delta)\right]\right\|_Y \leq \Delta\frac{\alpha_2}{\alpha_1}\left\|\underline{\mathbf{u}}^{(P-1)} - \underline{\mathbf{u}}_\Delta\right\|_X.$$

And thus 2.4 becomes

$$\left\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}_\Delta\right\|_X \leq \Delta\frac{\alpha_2}{\alpha_1}\left\|\underline{\mathbf{u}}^{(P-1)} - \underline{\mathbf{u}}_\Delta\right\|_X. \quad (2.5)$$

By repeating these calculations recursively we get the thesis. □

The result is very general and may seem a little abstract but let's notice that if the operator $\mathcal{L}^1_\Delta$ is "easy" to solve (as it will be in the applications) then 2.1 represents a simple receipt to approximate arbitrarily well the solution $\underline{\mathbf{u}}_\Delta$ of $\mathcal{L}^2_\Delta$. The convergence to the solution for $P \to +\infty$ is ensured independently of the starting vector $\mathbf{u}^{(0)}$ provided that $\Delta\frac{\alpha_2}{\alpha_1} < 1$. Let's make some final observations.

Observation 2. *If the solution $\underline{\mathbf{u}}_\Delta$ of $\mathcal{L}^2_\Delta$ is a R-th order approximation of the exact solution $\underline{\mathbf{u}}^{ex}$ of a more general problem then a R-th order approximation $\tilde{\underline{\mathbf{u}}}$ of $\underline{\mathbf{u}}_\Delta$ is a R-th order approximation of the exact solution $\underline{\mathbf{u}}^{ex}$ as well. It follows easily by applying the triangular inequality*

$$\|\tilde{\underline{\mathbf{u}}} - \underline{\mathbf{u}}^{ex}\|_X \leq \|\tilde{\underline{\mathbf{u}}} - \underline{\mathbf{u}}_\Delta\|_X + \|\underline{\mathbf{u}}_\Delta - \underline{\mathbf{u}}^{ex}\|_X \leq O(\Delta^{R+1}) + O(\Delta^{R+1}) = O(\Delta^{R+1}).$$

Thus in this case we have to use the Deferred Correction procedure to approximate $\underline{\mathbf{u}}_\Delta$ exactly with R-th order accuracy, any other extra precision in approximating $\underline{\mathbf{u}}_\Delta$ would be useless[3]*. We will be more precise about the accuracy and the number of iterations needed to reach the highest possible accuracy later in the applications.*

Observation 3. *The independecy of the coefficients α_1 and α_2 on Δ may naively lead us to think that we can easily consider the limit for $\Delta \to 0$ but this is not true at least from a theoretical point of view. In fact the whole framework has a non-trivial dependence on Δ: the operators $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$ as well as the spaces X and Y and the norms on them depend on Δ. Despite this the numerical experiments seem to confirm the desired order of convergence.*

Now that we have introduced and proved the Deferred Correction in a general framework, the only thing that it is left to do to apply it in a more specific context is to characterize the objects needed (the operators $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$, the spaces X and Y, the norms) and verify that the three hypoteses are satisfied.

[3]In the context of the integration of ordinary differential equations, the solution $\underline{\mathbf{u}}_\Delta$ of $\mathcal{L}^2_\Delta$ will be a $(M+1)$-order accurate approximation of the exact solution $\underline{\mathbf{u}}^{ex}$ and we will perform $P = M+1$ iterations to get $\underline{\mathbf{u}}^{(M+1)}$ which will be a $(M+1)$-order accurate approximation of $\underline{\mathbf{u}}_\Delta$ and thus of $\underline{\mathbf{u}}^{ex}$. Also in the application to the numerical methods for hyperbolic systems of balance laws we will have an analogous situation.

Chapter 3

The Deferred Correction for systems of ODEs

We will now apply the Deferred Correction method to the context of the integration of systems of ordinary differential equations. Let's consider the Cauchy problem

$$\begin{cases} \frac{d}{dt}\mathbf{u}(t) = \mathbf{G}(t, \mathbf{u}(t)) \\ \mathbf{u}(t_0) = \mathbf{u}_0 \end{cases} \tag{3.1}$$

with $\mathbf{u} \in \mathbb{R}^N$ and the usual hypoteses of regularity which ensure the existence of a unique solution i.e. $\mathbf{G}$ continuous and Lipschitz-continuous with respect to $\mathbf{u}$ uniformly with respect to t with a Lipschitz constant L. We would like to approximate numerically the solution at the time $t_0 + \Delta t$. We are now going to put the problem in a Deferred Correction framework. Clearly in this case the parameter Δ is the time step Δt. Let's now define the operators $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$ and the normed vector spaces X and Y.

3.1 Derivation of the operators $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$

As anticipated in the abstract context $\mathcal{L}^2_\Delta$ is a "tough" high-order implicit operator which provides the desired accuracy but that is difficult to solve, instead $\mathcal{L}^1_\Delta$ is a "naive" low-order explicit operator that is easy to solve but that is not enough accurate. When we write about the accuracy of these operators we mean the approximation accuracy of their solutions with respect to the (exact) solution of the general problem that we would like to solve which is in this case the system of ODEs 3.1.

3.1.1 Derivation of $\mathcal{L}^2_\Delta$

We are interested in a high-order approximation of the solution to the Cauchy problem 3.1 at the time $t_0 + \Delta t$. In order to get it we introduce a set of nodes in the interval $[t_0, t_0 + \Delta t]$ where we will consider the approximations of the values of the solution to our system of ODEs. The first node coincides with t_0, the last one with $t_0 + \Delta t$. Clearly

we are interested just in the approximation of the solution in the last node but we will need also the approximations of the solution in the other nodes.

We consider the interval $[t_0, t_0 + \Delta t]$ and define $M+1$ nodes t^m with $m = 0, 1, \ldots, M$ such that

$$t_0 = t^0 < t^1 < \cdots < t^M = t_0 + \Delta t$$

like in figure 3.1.

Observation 4. *We can assume the nodes to be equispaced for simplicity's sake but this is not mandatory and the procedure would not change for any other arbitrary distribution of them. Anyway if we assume an equispaced distribution, the distance between two consecutive points is* $\frac{\Delta t}{M}$.

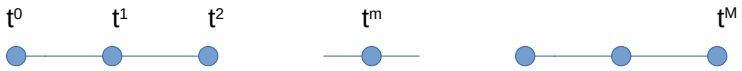


Figure 3.1: Nodes in the interval $[t^0, t^0 + \Delta t]$.

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 $\mathbf{u}^0 = \mathbf{u}(t^0) = \mathbf{u}(t_0) = \mathbf{u}_0$ without any approximation, but obviously the exact solutions as well as the approximations of solution in the other nodes are unkonwns.

An exact integration of the system of ODEs would result in

$$\mathbf{u}(t^m) - \mathbf{u}^0 - \int_{t^0}^{t^m} \mathbf{G}(t, \mathbf{u}(t))dt = \mathbf{0} \tag{3.2}$$

from which we would have the exact solution $\mathbf{u}(t^m)$. Unfortunately we cannot perform in general the exact integration (also because we actually do not know $\mathbf{u}$ in $[t^0, t^m]$) and we need to make some approximations. Let's approximate $\mathbf{G}(t, \mathbf{u}(t))$ with M-order

accuracy through the Lagrange polynomials of degree M corresponding to the $M+1$ nodes t^m with $m = 0, 1, \ldots, M$ i.e.

$$\mathbf{G}(t, \mathbf{u}(t)) = \sum_{l=0}^{M} \mathbf{G}(t^l, \mathbf{u}(t^l))\psi^l(t) + O(\Delta t^{M+1}) \tag{3.3}$$

with

$$\psi^l(t) = \prod_{\substack{m=0 \\ m \neq l}}^{M} \frac{(t - t^m)}{(t^l - t^m)} \quad \forall l = 0, 1, \ldots, M.$$

In figure 3.2 are shown the Lagrange polynomials corresponding to equispaced nodes in the reference interval $[0, 1]$ from order 1 to 6.

Thus if we substitute the approximation 3.3 in the formula got by the exact integration 3.2 we get a relation involving $\mathbf{u}^m$ which is a $(M+1)$-order accurate approximation of $\mathbf{u}(t^m)$

$$\mathbf{u}^m - \mathbf{u}^0 - \int_{t^0}^{t^m} \sum_{l=0}^{M} \mathbf{G}(t^l, \mathbf{u}(t^l))\psi^l(t)dt = \mathbf{0}. \tag{3.4}$$

Observation 5. *We cannot use this relation to calculate $\mathbf{u}^m$ because we do not know the values $\mathbf{u}(t^l)$. So we have not finished yet and 3.4 is not yet our final operator $\mathcal{L}^2_\Delta$. We need to modify it a bit.*

Before modifying 3.4 to get something that we can actually "use", let's verify the claim on its $(M+1)$-order accuracy.

Proposition 1. *$\mathbf{u}^m$ satisfying 3.4 is a $(M+1)$-order accurate approximation of $\mathbf{u}(t^m)$.*

Proof. Let's compute $\mathbf{u}(t^m) - \mathbf{u}^m$ with $\mathbf{u}^m$ got by 3.4. From 3.2, 3.4 and the M-order accuracy on the approximation of $\mathbf{G}(t, \mathbf{u}(t))$ expressed in equation 3.3 we have

$$\mathbf{u}(t^m) - \mathbf{u}^m = \mathbf{u}^0 + \int_{t^0}^{t^m} \mathbf{G}(t, \mathbf{u}(t))dt - \mathbf{u}^0 - \int_{t^0}^{t^m} \sum_{l=0}^{M} \mathbf{G}(t^l, \mathbf{u}(t^l))\psi^l(t)dt =$$

$$= \int_{t^0}^{t^m} \left[\mathbf{G}(t, \mathbf{u}(t)) - \sum_{l=0}^{M} \mathbf{G}(t^l, \mathbf{u}(t^l))\psi^l(t) \right] dt = \int_{t^0}^{t^m} O(\Delta t^{M+1})dt = O(\Delta t^{M+2}).$$

□

As we said before we do not have the values $\mathbf{u}(t^l)$ and we need to modify 3.4 to get somthing which can actually be used to get $\mathbf{u}^m$.

The idea is to use the approximated values $\mathbf{u}^l$ in place of them, thus getting an implicit formulation. By doing this we do not lose any accuracy as we will see in a few lines. The implicit formulation that allows us to get $\mathbf{u}^m$ with $M+1$-order accuracy reads then

$$\mathbf{u}^m - \mathbf{u}^0 - \int_{t^0}^{t^m} \sum_{l=0}^{M} \mathbf{G}(t^l, \mathbf{u}^l)\psi^l(t)dt = \mathbf{0} \quad \forall m = 1, 2, \ldots, M \tag{3.5}$$

This is the operator $\mathcal{L}^2_\Delta$ that we were looking for but before going ahead and rearrange it in a more clever way let's verify that, despite the extra approximation introduced with respect to 3.4, the order of accuracy is still $M+1$.

Proposition 2. $\mathbf{u}^m$ *satisfying 3.5 is a* $(M+1)$*-order accurate approximation of* $\mathbf{u}(t^m)$.

Proof. Let's first show that, by substituiting $\mathbf{u}^l$ to $\mathbf{u}(t^l)$, we are making an $O(\Delta t^{M+2})$ approximation on $\sum_{l=0}^{M} \mathbf{G}(t^l, \mathbf{u}(t^l))\psi^l(t)$. By applying the triangular inequality and remembering that the Lagrangian functions are bounded in absolute value by a positive constant[1], say C, we get

$$\left\| \sum_{l=0}^{M} \left[\mathbf{G}(t^l, \mathbf{u}^l) - \mathbf{G}(t^l, \mathbf{u}(t^l)) \right] \psi^l(t) \right\|_\infty \leq C \sum_{l=0}^{M} \left\| \mathbf{G}(t^l, \mathbf{u}^l) - \mathbf{G}(t^l, \mathbf{u}(t^l)) \right\|_\infty$$

and due to the Lipschitz-continuity of the flux with respect to $\mathbf{u}$ uniformly with respect to t with Lispchitz constant L and the fact that $\mathbf{u}^l$ is a $(M+1)$-order approximation of $\mathbf{u}(t^l)$ we get

$$C \sum_{l=0}^{M} \left\| \mathbf{G}(t^l, \mathbf{u}^l) - \mathbf{G}(t^l, \mathbf{u}(t^l)) \right\|_\infty \leq CL \sum_{l=0}^{M} \left\| \mathbf{u}^l - \mathbf{u}(t^l) \right\|_\infty \leq CL \sum_{l=0}^{M} O(\Delta t^{M+2}) = O(\Delta t^{M+2}).$$

In the end we get

$$\sum_{l=0}^{M} \left[\mathbf{G}(t^l, \mathbf{u}^l) - \mathbf{G}(t^l, \mathbf{u}(t^l)) \right] \psi^l(t) = O(\Delta t^{M+2}) \tag{3.6}$$

Thus by substituting $\mathbf{u}^l$ to $\mathbf{u}(t^l)$ we are making an $O(\Delta t^{M+2})$ approximation in the evaluation of $G(t, \mathbf{u}(t))$ which doesn't affect the accuracy. In fact if we consider $\mathbf{u}(t^m) - \mathbf{u}^m$ with $\mathbf{u}^m$ got by equation 3.5, remembering 3.2, we have

$$\mathbf{u}(t^m) - \mathbf{u}^m = \mathbf{u}^0 + \int_{t^0}^{t^m} \mathbf{G}(t, \mathbf{u}(t))dt - \mathbf{u}^0 - \int_{t^0}^{t^m} \sum_{l=0}^{M} \mathbf{G}(t^l, \mathbf{u}^l)\psi^l(t)dt =$$

$$= \int_{t^0}^{t^m} \left[\mathbf{G}(t, \mathbf{u}(t)) - \sum_{l=0}^{M} \mathbf{G}(t^l, \mathbf{u}^l)\psi^l(t) \right] dt.$$

Recalling now 3.6 we have

$$\int_{t^0}^{t^m} \left[\mathbf{G}(t, \mathbf{u}(t)) - \sum_{l=0}^{M} \mathbf{G}(t^l, \mathbf{u}^l)\psi^l(t) \right] dt = \int_{t^0}^{t^m} \left[\mathbf{G}(t, \mathbf{u}(t)) - \sum_{l=0}^{M} \mathbf{G}(t^l, \mathbf{u}(t^l))\psi^l(t) + O(\Delta t^{M+2}) \right] dt$$

[1] Once that the nodes are fixed the Lagrangian functions are continuous functions over a compact set and for the Weierstrass theorem they are bounded.

and by using again 3.3 we have our desired result on the accuracy

$$\int_{t^0}^{t^m} \left[\mathbf{G}(t, \mathbf{u}(t)) - \sum_{l=0}^{M} \mathbf{G}(t^l, \mathbf{u}(t^l))\psi^l(t) + O(\Delta t^{M+2}) \right] dt = \int_{t^0}^{t^m} \left[O(\Delta t^{M+1}) + O(\Delta t^{M+2}) \right] dt =$$

$$= O(\Delta t^{M+2}) + O(\Delta t^{M+3}) = O(\Delta t^{M+2}).$$

□

Coming back to our initial goal, we will now define the operator $\mathcal{L}_\Delta^2$ directly from 3.5 which is recalled for clarity

$$\mathbf{u}^m - \mathbf{u}^0 - \int_{t^0}^{t^m} \sum_{l=0}^{M} \mathbf{G}(t^l, \mathbf{u}^l)\psi^l(t)dt = \mathbf{0} \quad \forall m = 1, 2, \ldots, M$$

Let's first recast it in a clever way by noticing that the vectors $\mathbf{G}(t^l, \mathbf{u}^l)$ do not depend on time so we can put the finite sum out of the integral as well as these vectors (even if, we remark, they are unknown because we don't know $\mathbf{u}^l$). We thus get

$$\mathbf{u}^m - \mathbf{u}^0 - \sum_{l=0}^{M} \mathbf{G}(t^l, \mathbf{u}^l) \int_{t^0}^{t^m} \psi^l(t)dt = \mathbf{0} \quad \forall m = 1, 2, \ldots, M$$

The Lagrangian polynomial functions $\psi^l(t)$ are known as well as the nodes t^m so we can perform exactly the integrals and set

$$\int_{t^0}^{t^m} \psi^l(t)dt = \Delta t \int_0^{\frac{t^m - t^0}{\Delta t}} \psi^l\left(\Delta t s + t^0\right) ds = \Delta t \theta_l^m \quad \forall m = 1, 2, \ldots, M \quad \forall l = 0, 1, \ldots, M.$$

where

$$\theta_l^m = \int_0^{\frac{t^m - t^0}{\Delta t}} \psi^l\left(\Delta t s + t^0\right) ds.$$

Notice that we made a change of variable which transformed the interval $[t^0, t^m]$ into $[0, \frac{t^m - t^0}{\Delta t}]$ or equivalently $[t^0, t^M]$ into $[0, 1]$. This allows us to work with normalized coefficients θ_l^m which depend just on the number and distribution of the nodes but not on Δt.

Observation 6. *In particular in the case of equispaced nodes we have*

$$\theta_l^m = \int_0^{\frac{m}{M}} \hat{\psi}^l(s)ds \quad \forall m = 1, 2, \ldots, M \quad \forall l = 0, 1, \ldots, M$$

where $\hat{\psi}^l \quad l = 0, 1, \ldots, M$ *are the Lagrangian functions associated to the* $M + 1$ *nodes* $0 < \frac{1}{M} < \frac{2}{M} < \cdots < \frac{M-1}{M} < 1$ *in the reference interval* $[0, 1]$.

So we have

$$\mathbf{u}^m - \mathbf{u}^0 - \Delta t \sum_{l=0}^{M} \theta_l^m \mathbf{G}(t^l, \mathbf{u}^l) = \mathbf{0} \quad \forall m = 1, 2, \ldots, M$$

through which we finally get our our implicit, high-order, "tough" operator $\mathcal{L}_\Delta^2 : \mathbb{R}^{(M\times N)} \to \mathbb{R}^{(M\times N)}$ defined as

$$\mathcal{L}_\Delta^2(\underline{\mathbf{u}}) = \begin{pmatrix} \mathbf{u}^M - \mathbf{u}^0 - \Delta t \sum_{l=0}^{M} \theta_l^M \mathbf{G}(t^l, \mathbf{u}^l) \\ \vdots \\ \mathbf{u}^m - \mathbf{u}^0 - \Delta t \sum_{l=0}^{M} \theta_l^m \mathbf{G}(t^l, \mathbf{u}^l) \\ \vdots \\ \mathbf{u}^1 - \mathbf{u}^0 - \Delta t \sum_{l=0}^{M} \theta_l^1 \mathbf{G}(t^l, \mathbf{u}^l) \end{pmatrix} \tag{3.7}$$

where

$$\underline{\mathbf{u}} = \begin{pmatrix} \mathbf{u}^M \\ \vdots \\ \mathbf{u}^m \\ \vdots \\ \mathbf{u}^1 \end{pmatrix}.$$

We equip $X = Y = \mathbb{R}^{(M\times N)}$ with the infinity norm $\|\cdot\|_\infty$. Let's make some observations to make things clearer.

Observation 7. *M is the number of nodes t^m in $[t_0, t_0 + \Delta t]$ in which we do not know the solution to the ODEs system i.e. all the nodes apart from t^0 (in this node we already know the exact solution $\mathbf{u}_0$). Instead N is the number of components of $\mathbf{u}$ i.e. the number of equations of our ODEs system.*

Observation 8. *It is very important to understand that in 3.7 we are making a little change of notation. The vector $\underline{\mathbf{u}}$ is the general argument of the operator $\mathcal{L}_\Delta^2$ and is obviously not in general its solution, it is just a not specified $(M \times N)$-dimesional vector. We are now referring to the vector $\mathbf{u}^m$ no more as the approximation of the solution in t^m but as N components of the general argument $\underline{\mathbf{u}}$ of $\mathcal{L}_\Delta^2$. We get our approximation of the solution if we solve the operator i.e. if find $\underline{\mathbf{u}}_\Delta$ s.t. $\mathcal{L}_\Delta^2(\underline{\mathbf{u}}_\Delta) = \mathbf{0}$. If we have*

$$\underline{\mathbf{u}}_\Delta = \begin{pmatrix} \mathbf{u}_\Delta^M \\ \vdots \\ \mathbf{u}_\Delta^m \\ \vdots \\ \mathbf{u}_\Delta^1 \end{pmatrix}$$

then we can say that $\mathbf{u}_\Delta^m$ is a $(M+1)$-order accurate approximation of the solution in t^m but in general $\mathbf{u}^m$ is just a not specified N-dimensional real vector.

Observation 9. *The operator $\mathcal{L}^2_\Delta$ is implicit and $(M+1)$-order accurate. Its solution, i.e. $\underline{\mathbf{u}}_\Delta$ s.t. $\mathcal{L}^2_\Delta(\underline{\mathbf{u}}_\Delta) = \mathbf{0}$, is made by all the solutions $\underline{\mathbf{u}}^m_\Delta$ to 3.5 which are $(M+1)$-order accurate approximations of the solution of the ODEs system in the nodes t^m $m = 1, 2, \ldots, M$.*

Observation 10. *$\mathbf{u}^0$ is not an unkonwn and for this reason it is not a argument of the operator $\mathcal{L}^2_\Delta$. It is "part" of the problem and it is "embedded" in the operator $\mathcal{L}^2_\Delta$.*

Observation 11. *The space $X = Y = \mathbb{R}^{(M\times N)}$ is finite dimensional so all the norms that we can define on it are equivalent. Despite this, the choice of the norms is crucial and not trivial. This is due to the fact that the constants α_1 and α_2 must not depend on Δ. If we choose wrong norms we may have that the sub-coercivity of $\mathcal{L}^1_\Delta$ is true but not the Lipschitz-continuity-like condition of $\mathcal{L}^1_\Delta - \mathcal{L}^2_\Delta$. In particular, as it will be soon clear, the first property is almost trivial to verify, while the second will need more attention to be proved. In general we have to take advantage of this fact by designing the norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ in such a way that the inequality of the sub-coercivity property of $\mathcal{L}^1_\Delta$ results in an equality. The basic idea is that we have to satisfy two "conflicting properties": roughly speaking, the sub-coercivity property of $\mathcal{L}^1_\Delta$ means that we need the norm $\|\cdot\|_X$ to be "smaller" than the norm $\|\cdot\|_Y$ while vice versa the Lipschitz-continuity-like condition of $\mathcal{L}^1_\Delta - \mathcal{L}^2_\Delta$ means the exact opposite. Again, since the first property is easier to be verified we need to design the norms to relax it at most, this results essential for the second property to be verified. This is the the general guideline to follow in the designing of two norms satisfying the desired properties if we wish to apply the Deferred Correction to any context. Anyway this will be clearer when we will consider the the application to the balance laws.*

Observation 12. *Let's remark that in principle we are not interested in all the components of $\mathcal{L}^2_\Delta$: we want a high-order approximation of the solution in the last node $t^M = t^0 + \Delta t$. Despite this, the construction of the operator is strictly based on the approximation of the solution also in the other nodes. Let's also notice that the accuracy of each component of the solution of $\mathcal{L}^2_\Delta$ with respect to the exact solution to our initial problem 3.1 in the corresponding node is $M+1$, not just for the component referred to the last node. Every component $\mathbf{u}^m_\Delta$ of the solution $\underline{\mathbf{u}}_\Delta$ is a $(M+1)$-order approximation of the exact solution to 3.1 in the correpsonding node t^m.*

3.1.2 Derivation of $\mathcal{L}^1_\Delta$

Also the operator $\mathcal{L}^1_\Delta$, just like $\mathcal{L}^2_\Delta$, will involve the approximations of the solution in the nodes t^m $m = 1, 2, \ldots, M$ and not just in the last one. The "naive" operator $\mathcal{L}^1_\Delta$ is obtained by simply applying the Euler method to solve the system of ODEs. In particular let's refer again to the exact integration 3.2 that we recall for clarity

$$\mathbf{u}(t^m) - \mathbf{u}^0 - \int_{t^0}^{t^m} \mathbf{G}(t, \mathbf{u}(t))dt = \mathbf{0} \quad m = 1, 2, \ldots, M.$$

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

$$\mathbf{u}^m - \mathbf{u}^0 - \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) = \mathbf{0} \tag{3.8}$$

where $\beta^m = \frac{t^m - t^0}{\Delta t}$. Just like for θ_l^m in the context of the $\mathcal{L}_\Delta^2$ operator, we put in evidence Δt to work with normalized coefficients. The Euler method is well known to provide a first order approximation of the exact solution.

Proposition 3. *The approximate solution $\mathbf{u}^m$ got through the Euler method 3.8 is first order accurate.*

Proof. Again we consider the difference between the exact solution $\mathbf{u}(t^m)$ to our ODEs system 3.1 and $\mathbf{u}^m$ got from 3.8. 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

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

because $\mathbf{u}^0 = \mathbf{u}(t^0) = \mathbf{u}(t_0) = \mathbf{u}_0$ and $\beta^m = \frac{t^m - t^0}{\Delta t}$. □

Directly from 3.8 we get our explicit, low-order, "naive" operator $\mathcal{L}_\Delta^2 : \mathbb{R}^{(M \times N)} \to \mathbb{R}^{(M \times N)}$ defined as

$$\mathcal{L}_\Delta^1(\underline{\mathbf{u}}) = \begin{pmatrix} \mathbf{u}^M - \mathbf{u}^0 - \Delta t \beta^M \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{u}^m - \mathbf{u}^0 - \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{u}^1 - \mathbf{u}^0 - \Delta t \beta^1 \mathbf{G}(t^0, \mathbf{u}^0) \end{pmatrix} \tag{3.9}$$

where

$$\underline{\mathbf{u}} = \begin{pmatrix} \mathbf{u}^M \\ \vdots \\ \mathbf{u}^m \\ \vdots \\ \mathbf{u}^1 \end{pmatrix}.$$

Let's remind that the norm chosen on $X = Y = \mathbb{R}^{(M \times N)}$ is the infinity norm $\|\cdot\|_\infty$. We can make observations analogue to the ones that we made for the operator $\mathcal{L}_\Delta^2$.

Observation 13. *In 3.9 the vector $\underline{\mathbf{u}}$ is the general argument of the operator $\mathcal{L}_\Delta^1$ and thus not in general its solution. We are now referring to the vector $\mathbf{u}^m$ as N components of the general argument $\underline{\mathbf{u}}$ of $\mathcal{L}_\Delta^1$ and not as the first order approximation of the exact solution in t^m that we would get by solving the operator.*

Observation 14. *The operator $\mathcal{L}_\Delta^1$ is explicit and easy to solve but just first-order accurate. Its solution is made by all the solutions to 3.8 which are first-order accurate approximations of the solution of the initial system od ODEs in the nodes t^m $m = 1, 2, \ldots, M$.*

Observation 15. *As in the previous case,* $\mathbf{u}^0$ *is not an unkonwn and for this reason it is not a argument of the operator* $\mathcal{L}^1_\Delta$*, it is a known vector embedded in the operator.*

Observation 16. *Also in this case let's remark that the solution to the operator* $\mathcal{L}^1_\Delta$ *would be first-order accurate in each of its components.*

3.2 From an operational perspective

Before proving that the operators $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$ fulfil the hypoteses required to apply the Deferred Correction let's make some practical considerations to give a clearer overview of the algorithm to implement. Assumed that the required properties hold, we can apply the method 2.1 in this specific context and get

$$\mathcal{L}^1_\Delta(\underline{\mathbf{u}}^{(p)}) = \mathcal{L}^1_\Delta(\underline{\mathbf{u}}^{(p-1)}) - \mathcal{L}^2_\Delta(\underline{\mathbf{u}}^{(p-1)}) \quad p = 1, 2, \dots, P \tag{3.10}$$

where

$$\underline{\mathbf{u}}^{(p)} = \begin{pmatrix} \mathbf{u}^{M,(p)} \\ \vdots \\ \mathbf{u}^{m,(p)} \\ \vdots \\ \mathbf{u}^{1,(p)} \end{pmatrix} \quad p = 1, 2, \dots, P$$

is the result of the p iteration which is made by M components $\mathbf{u}^{m,(p)}$ corresponding to the approximations of the solution in the subtimesteps t^m $m = 1, 2, \dots, M$. Each one of them is itself a vector with N components where N is the number of equations of the system of ODEs 3.1. In order to make things clearer let's look at the grid in figure 3.3. On the ordinate axis we have the iterations while on the abscissa axis we have the subtimesteps. The procedure 3.10 results in an explicit iterative algorithm due to the fact that the operator $\mathcal{L}^1_\Delta$ is explicit. In fact the generic p iteration reads

$$\begin{pmatrix} \vdots \\ \mathbf{u}^{m,(p)} - \mathbf{u}^0 - \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathbf{u}^{m,(p-1)} - \mathbf{u}^0 - \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \end{pmatrix} +$$

$$- \begin{pmatrix} \vdots \\ \mathbf{u}^{m,(p-1)} - \mathbf{u}^0 - \Delta t \sum_{l=0}^{M} \theta_l^m \mathbf{G}(t^l, \mathbf{u}^{l,(p-1)}) \\ \vdots \end{pmatrix}$$

and we can calculate $\underline{\mathbf{u}}^{(p)}$ in an explicit way

$$\underline{\mathbf{u}}^{(p)} = \begin{pmatrix} \vdots \\ \mathbf{u}^{m,(p)} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathbf{u}^0 + \Delta t \sum_{l=0}^{M} \theta_l^m \mathbf{G}(t^l, \mathbf{u}^{l,(p-1)}) \\ \vdots \end{pmatrix}$$

Observation 17. *We need a starting vector* $\underline{\mathbf{u}}^{(0)}$ *for our iteration process. So we assume*

$$\underline{\mathbf{u}}^{(0)} = \begin{pmatrix} \mathbf{u}^0 \\ \vdots \\ \mathbf{u}^0 \\ \vdots \\ \mathbf{u}^0 \end{pmatrix}$$

i.e. we assume for every component $m = 1, 2, \ldots, M$ *of the argument,* $\mathbf{u}^{m,(0)} = \mathbf{u}^0$ *which is known.*

Observation 18. *When we write* $\mathbf{u}^{0,(p)}$ *in order to evaluate* $\sum_{l=0}^{M} \theta_l^m \mathbf{G}(t^l, \mathbf{u}^{l,(p-1)})$ *for* $m = 1, 2, \ldots, M$ *we clearly mean* $\mathbf{u}^0$*. As we already underlined in the observations 10 and 15, the components* $\mathbf{u}^m$ *of the global argument* $\underline{\mathbf{u}}^m$ *of* $\mathcal{L}_\Delta^1$ *and* $\mathcal{L}_\Delta^2$ *correspond to the approximate solutions to the system of ODEs 3.1 in the nodes* t^m $m = 1, 2, \ldots, M$ *where the solution is unkonwn. Instead in* t^0 *we have* $\mathbf{u}^0 = \mathbf{u}(t^0) = \mathbf{u}(t_0) = \mathbf{u}_0$*. The vector* $\mathbf{u}^0$ *is fixed, it is part of the initial probem, it's not a variable. We should in principle write*

$$\sum_{l=0}^{M} \theta_l^m \mathbf{G}(t^l, \mathbf{u}^{l,(p-1)}) = \theta_0^m \mathbf{G}(t^0, \mathbf{u}^0) + \sum_{l=1}^{M} \theta_l^m \mathbf{G}(t^l, \mathbf{u}^{l,(p-1)})$$

but to keep the notation more compact we avoid this and we set

$$\mathbf{u}^{0,(p)} = \mathbf{u}^0 \quad p = 0, 1, \ldots, P.$$

In practice we are making an abuse of notation: we are using the notation through which we refer in general to the variables $\mathbf{u}^{m,(p)}$ *to refer to the constant vector* $\mathbf{u}^0$ *whenever* m *or* p *are equal to* 0*.*

Now the algorithm is clear and we can focus on the optimal number of iterations to get the highest possible accuracy.

As already anticipated, at least in part, in observation 2 the Deferred Correction doesn't provide "directly" an approximation of the exact solution (in the nodes t^m with $m = 1, 2, \ldots, M$) to our ODEs system but an approximation of the solution to $\mathcal{L}_\Delta^2$ which is itself a $(M+1)$-order approximation of the exact solution. And it would be useless to make a "too high" number of iterations to approximate it with an accuracy order higher than $M + 1$. To be clearer, if we define the vector

$$\underline{\mathbf{u}}^{ex} = \begin{pmatrix} \mathbf{u}(t^M) \\ \vdots \\ \mathbf{u}(t^m) \\ \vdots \\ \mathbf{u}(t^1) \end{pmatrix}$$

made by the exact solution to our ODEs system evaluated in the nodes t^m $m = 1, 2, \ldots, M$, the Deferred Correction does not approximate directly it but $\underline{\mathbf{u}}_\Delta$ which is the solution to $\mathcal{L}^2_\Delta$ and a $(M+1)$-order accurate approximation of $\underline{\mathbf{u}}^{ex}$. This fact is shown in the next proposition.

Proposition 4. *The vector $\underline{\mathbf{u}}^{(P)}$ got from the final iteration P of the Deferred Correction algorithm 3.10 is a P-order approximation of $\underline{\mathbf{u}}_\Delta$, the solution of the operator $\mathcal{L}^2_\Delta$.*

Proof. From our first abstract accuracy estimate 2.2 we have in this case

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

We just suffice to prove that $\left\|\underline{\mathbf{u}}^{(0)} - \underline{\mathbf{u}}_\Delta\right\|_\infty$ is $O(\Delta t)$. To do this let's consider $\tilde{\underline{\mathbf{u}}}_\Delta$ solution to $\mathcal{L}^1_\Delta$ i.e. such that $\mathcal{L}^1_\Delta(\tilde{\underline{\mathbf{u}}}_\Delta) = \mathbf{0}$. As we pointed out several times $\mathcal{L}^1_\Delta$ is a first-order accurate operator in the sense that its solution is first-order accurate with respect to the the vector $\underline{\mathbf{u}}^{ex}$ containing the evaluations of the exact solution in the nodes t^m $m = 1, 2, \ldots, M$. Moreover it is explicit and very easy to solve. Directly from its definition 3.9, by solving $\mathcal{L}^1_\Delta(\tilde{\underline{\mathbf{u}}}_\Delta) = \mathbf{0}$, we have

$$\tilde{\underline{\mathbf{u}}}_\Delta = \begin{pmatrix} \tilde{\mathbf{u}}^M_\Delta \\ \vdots \\ \tilde{\mathbf{u}}^m_\Delta \\ \vdots \\ \tilde{\mathbf{u}}^1_\Delta \end{pmatrix} = \begin{pmatrix} \mathbf{u}^0 + \Delta t \beta^M \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{u}^0 + \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{u}^0 + \Delta t \beta^1 \mathbf{G}(t^0, \mathbf{u}^0) \end{pmatrix}.$$

Instead the operator $\mathcal{L}^2_\Delta$ is $(M+1)$-order accurate since its solution $\underline{\mathbf{u}}_\Delta$ is $(M+1)$-order accurate with respect to $\underline{\mathbf{u}}^{ex}$. Thus, by adding and subtracting $\underline{\mathbf{u}}^{ex}$ and $\tilde{\underline{\mathbf{u}}}_\Delta$ and by applying the triangular inequality we have

$$\left\|\underline{\mathbf{u}}^{(0)} - \underline{\mathbf{u}}_\Delta\right\|_\infty \leq \left\|\underline{\mathbf{u}}^{(0)} - \tilde{\underline{\mathbf{u}}}_\Delta\right\|_\infty + \|\tilde{\underline{\mathbf{u}}}_\Delta - \underline{\mathbf{u}}^{ex}\|_\infty + \|\underline{\mathbf{u}}^{ex} - \underline{\mathbf{u}}_\Delta\|_\infty =$$

$$= \left\|\underline{\mathbf{u}}^{(0)} - \tilde{\underline{\mathbf{u}}}_\Delta\right\|_\infty + O(\Delta t^2) + O(\Delta t^{M+2}).$$

From a direct computation we have $\left\|\underline{\mathbf{u}}^{(0)} - \tilde{\underline{\mathbf{u}}}_\Delta\right\| = O(\Delta t)$ in fact

$$\tilde{\underline{\mathbf{u}}}_\Delta - \underline{\mathbf{u}}^{(0)} = \begin{pmatrix} \mathbf{u}^0 + \Delta t \beta^M \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{u}^0 + \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{u}^0 + \Delta t \beta^1 \mathbf{G}(t^0, \mathbf{u}^0) \end{pmatrix} - \begin{pmatrix} \mathbf{u}^0 \\ \vdots \\ \mathbf{u}^0 \\ \vdots \\ \mathbf{u}^0 \end{pmatrix} = \Delta t \begin{pmatrix} \beta^M \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \beta^m \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \beta^1 \mathbf{G}(t^0, \mathbf{u}^0) \end{pmatrix} = O(\Delta t).$$

Remember that the coefficients β^m are normalized coefficients that depend just on the distribution of the nodes t^m but not on Δt and $\mathbf{G}(t^0, \mathbf{u}^0)$ is a known constant vector. It follows that $\left\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}_\Delta\right\|_\infty = O(\Delta t^{P+1})$ which is the thesis. □

Thus we have that $\underline{\mathbf{u}}^{(P)}$, resulting from the Deferred Correction procedure with P iterations, is a P-order accurate approximation of the solution $\underline{\mathbf{u}}_\Delta$ to $\mathcal{L}^2_\Delta$ which is itself a $(M+1)$-order approximation of the vector $\underline{\mathbf{u}}^{ex}$ made by the evaluations of the exact solution in the nodes $m = 1, 2, \ldots, M$. From this we immediately see that the highest possible accuracy with respect to $\underline{\mathbf{u}}^{ex}$ is bounded by $M+1$ and it is reached with a number of iterations at least equal to $M+1$.

Proposition 5. *The accuracy of $\underline{\mathbf{u}}^{(P)}$, resulting from the Deferred Correction procedure with P iterations, with respect to the vector $\underline{\mathbf{u}}^{ex}$ containing the evaluations of the exact solution in the nodes $m = 1, 2, \ldots, M$ is $\min\{P, M+1\}$.*

Proof. Let's consider $\left\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}^{ex}\right\|_\infty$. We add and subtract $\underline{\mathbf{u}}_\Delta$, solution of $\mathcal{L}^2_\Delta$, and use the trianguar inequality, thus we have

$$\left\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}^{ex}\right\|_\infty \leq \left\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}_\Delta\right\|_\infty + \left\|\underline{\mathbf{u}}_\Delta - \underline{\mathbf{u}}^{ex}\right\|_\infty .$$

As we repeated several times, since the operator $\mathcal{L}^2_\Delta$ is $(M+1)$-order accurate we have

$$\left\|\underline{\mathbf{u}}_\Delta - \underline{\mathbf{u}}^{ex}\right\|_\infty = O(\Delta t^{M+2})$$

and since, due to the proposition 4, $\underline{\mathbf{u}}^{(P)}$ is a P-order approximation of $\underline{\mathbf{u}}_\Delta$ it holds

$$\left\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}_\Delta\right\|_\infty = O(\Delta t^{P+1}).$$

Thus we can write

$$\left\|\underline{\mathbf{u}}^{(P)} - \underline{\mathbf{u}}^{ex}\right\|_\infty \leq O(\Delta t^{M+2}) + O(\Delta t^{P+1})$$

which is the thesis. □

Observation 19. *If we increase P we increase the accuracy of $\underline{\mathbf{u}}^{(P)}$ with respect to $\underline{\mathbf{u}}_\Delta$ which is a $(M+1)$-order approximation of $\underline{\mathbf{u}}^{ex}$. Since our aim is approximating $\underline{\mathbf{u}}^{ex}$ and not $\underline{\mathbf{u}}_\Delta$, increasing the number of iterations is useful only until we reach $(M+1)$-order accuracy. The optimal number of iterations is thus $P = M+1$, any other extra iteration would just be a waste of computational resources so we set $P = M+1$. Anyway to be more general, we will continue to use P to refer to the number of iterations keeping in mind that the accuracy of $\underline{\mathbf{u}}^{(P)}$ with respect to $\underline{\mathbf{u}}^{ex}$ is $\min\{P, M+1\}$ and that the most convenient number of iterations is $P = M+1$ which is the small number of iterations to provide $M+1$ accuracy.*

Observation 20. *Despite the esteem on the accuracy of $\underline{\mathbf{u}}^{(P)}$ being uniform with respect to all its components we are just interested in the component associated with the node M becasuse our aim is to have an approximation of the solution to the system of ODEs 3.1 in $t^M = t + \Delta t$. So at the end we take $\mathbf{u}^{M,(P)}$ as our desired approximated solution at the time $t + \Delta t$.*

Now we can finally prove that the hypoteses on the operators $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$ needed to apply the Deferred Correction method are fulfilled.

3.3 Proof of the properties on $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$

Let's recall the hypoteses 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 $\mathcal{L}^2_\Delta$**
$\exists! \underline{\mathbf{u}}_\Delta \in \mathbb{R}^{(M\times N)}$ solution of $\mathcal{L}^2_\Delta$ i.e. such that $\mathcal{L}^2_\Delta(\underline{\mathbf{u}}_\Delta) = \mathbf{0}$;

ii) **Sub-coercivity of $\mathcal{L}^1_\Delta$**
$\exists \alpha_1 \geq 0$ independent of Δt s.t.

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

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

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

The first property i.e. the exsistence of a unique solution to $\mathcal{L}^2_\Delta$ is assumed. Let's now consider two generic vectors $\underline{\mathbf{v}}, \underline{\mathbf{w}} \in \mathbb{R}^{(M\times N)}$

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

with $\mathbf{v}^m$ and $\mathbf{w}^m$ $m = 1, 2, \ldots, M$ generic N-dimensional vectors and prove the other two conditions. As anticipated in the observation 11, the sub-coercivity of $\mathcal{L}^1_\Delta$ will be straightforward instead the Lipschitz-continuity-like condition of $\mathcal{L}^1_\Delta - \mathcal{L}^2_\Delta$ will be less trivial (even more delicate in the application of the Deferred Correction to the balance laws).

3.3.1 Sub-coercivity of $\mathcal{L}^1_\Delta$

From a direct computation we have

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

so we have $\mathcal{L}^1_\Delta(\underline{\mathbf{v}}) - \mathcal{L}^1_\Delta(\underline{\mathbf{w}}) = \underline{\mathbf{v}} - \underline{\mathbf{w}}$ and then

$$\left\|\mathcal{L}^1_\Delta(\underline{\mathbf{v}}) - \mathcal{L}^1_\Delta(\underline{\mathbf{w}})\right\|_\infty = \left\|\underline{\mathbf{v}} - \underline{\mathbf{w}}\right\|_\infty$$

and thus the sub-coercivity of $\mathcal{L}^1_\Delta$ is verified and results in an equality. Again we remark that $\mathbf{u}^0$ is given, it's part of the problem and embedded in the operators $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$.

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

Again we consider a direct computation

$$\left[\mathcal{L}^1_\Delta(\underline{\mathbf{v}}) - \mathcal{L}^2_\Delta(\underline{\mathbf{v}})\right] - \left[\mathcal{L}^1_\Delta(\underline{\mathbf{w}}) - \mathcal{L}^2_\Delta(\underline{\mathbf{w}})\right] =$$

$$= \begin{bmatrix} \begin{pmatrix} \mathbf{v}^M - \mathbf{u}^0 - \Delta t \beta^M \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{v}^m - \mathbf{u}^0 - \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{v}^1 - \mathbf{u}^0 - \Delta t \beta^1 \mathbf{G}(t^0, \mathbf{u}^0) \end{pmatrix} - \begin{pmatrix} \mathbf{v}^M - \mathbf{u}^0 - \Delta t \sum_{l=0}^M \theta_l^M \mathbf{G}(t^l, \mathbf{v}^l) \\ \vdots \\ \mathbf{v}^m - \mathbf{u}^0 - \Delta t \sum_{l=0}^M \theta_l^m \mathbf{G}(t^l, \mathbf{v}^l) \\ \vdots \\ \mathbf{v}^1 - \mathbf{u}^0 - \Delta t \sum_{l=0}^M \theta_l^1 \mathbf{G}(t^l, \mathbf{v}^l) \end{pmatrix} \end{bmatrix} +$$

$$- \begin{bmatrix} \begin{pmatrix} \mathbf{w}^M - \mathbf{u}^0 - \Delta t \beta^M \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{w}^m - \mathbf{u}^0 - \Delta t \beta^m \mathbf{G}(t^0, \mathbf{u}^0) \\ \vdots \\ \mathbf{w}^1 - \mathbf{u}^0 - \Delta t \beta^1 \mathbf{G}(t^0, \mathbf{u}^0) \end{pmatrix} - \begin{pmatrix} \mathbf{w}^M - \mathbf{u}^0 - \Delta t \sum_{l=0}^M \theta_l^M \mathbf{G}(t^l, \mathbf{w}^l) \\ \vdots \\ \mathbf{w}^m - \mathbf{u}^0 - \Delta t \sum_{l=0}^M \theta_l^m \mathbf{G}(t^l, \mathbf{w}^l) \\ \vdots \\ \mathbf{w}^1 - \mathbf{u}^0 - \Delta t \sum_{l=0}^M \theta_l^1 \mathbf{G}(t^l, \mathbf{w}^l) \end{pmatrix} \end{bmatrix} \tag{3.11}$$

where clearly $\mathbf{v}^0 = \mathbf{w}^0 = \mathbf{u}^0$. As we pointed out several times expecially in the derivation of the operators $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$, for example in the observations 10 and 15, $\mathbf{u}^0$ 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 as we did for $\mathbf{u}^{0,(p)}$. The reader is referred to the observation 18 for more clarity. By elementary algebra 3.11 becomes

$$\left[\mathcal{L}^1_\Delta(\underline{\mathbf{v}}) - \mathcal{L}^2_\Delta(\underline{\mathbf{v}})\right] - \left[\mathcal{L}^1_\Delta(\underline{\mathbf{w}}) - \mathcal{L}^2_\Delta(\underline{\mathbf{w}})\right] = \Delta t \sum_{l=0}^M \begin{pmatrix} \theta_l^M \left[\mathbf{G}(t^l, \mathbf{v}^l) - \mathbf{G}(t^l, \mathbf{w}^l)\right] \\ \vdots \\ \theta_l^m \left[\mathbf{G}(t^l, \mathbf{v}^l) - \mathbf{G}(t^l, \mathbf{w}^l)\right] \\ \vdots \\ \theta_l^1 \left[\mathbf{G}(t^l, \mathbf{v}^l) - \mathbf{G}(t^l, \mathbf{w}^l)\right] \end{pmatrix}.$$

By applying the triangular inequality and remembering that θ_l^m $m = 1, 2, \ldots, M$ $l = 0, 1, \ldots, M$ are fixed normalized constant coefficients not depending on Δt, thus bounded in absolute value by a positive constant C, 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 we have

$$\left\| \left[\mathcal{L}^1_\Delta(\underline{\mathbf{v}}) - \mathcal{L}^2_\Delta(\underline{\mathbf{v}})\right] - \left[\mathcal{L}^1_\Delta(\underline{\mathbf{w}}) - \mathcal{L}^2_\Delta(\underline{\mathbf{w}})\right] \right\|_\infty = \Delta t \left\| \sum_{l=0}^M \begin{pmatrix} \theta_l^M \left[\mathbf{G}(t^l, \mathbf{v}^l) - \mathbf{G}(t^l, \mathbf{w}^l)\right] \\ \vdots \\ \theta_l^m \left[\mathbf{G}(t^l, \mathbf{v}^l) - \mathbf{G}(t^l, \mathbf{w}^l)\right] \\ \vdots \\ \theta_l^1 \left[\mathbf{G}(t^l, \mathbf{v}^l) - \mathbf{G}(t^l, \mathbf{w}^l)\right] \end{pmatrix} \right\|_\infty \leq$$

$$\leq \Delta t C \sum_{l=0}^{M} \left\| \begin{pmatrix} \mathbf{G}(t^l, \mathbf{v}^l) - \mathbf{G}(t^l, \mathbf{w}^l) \\ \vdots \\ \mathbf{G}(t^l, \mathbf{v}^l) - \mathbf{G}(t^l, \mathbf{w}^l) \\ \vdots \\ \mathbf{G}(t^l, \mathbf{v}^l) - \mathbf{G}(t^l, \mathbf{w}^l) \end{pmatrix} \right\|_\infty = \Delta t C \sum_{l=0}^{M} \left\| \mathbf{G}(t^l, \mathbf{v}^l) - \mathbf{G}(t^l, \mathbf{w}^l) \right\|_{\infty,N} \leq$$

$$\leq \Delta t C \sum_{l=0}^{M} L \left\| \mathbf{v}^l - \mathbf{w}^l \right\|_{\infty,N} \leq \Delta t C L M \left\| \underline{\mathbf{v}} - \underline{\mathbf{w}} \right\|_\infty$$

where C is a positive constant depending on the coefficients θ_l^m and not on Δt. The last inequality follows from the fact that $\underline{\mathbf{v}} - \underline{\mathbf{w}}$ contains as components all the vectors $\mathbf{v}^l - \mathbf{w}^l$ and thus

$$\left\| \mathbf{v}^l - \mathbf{w}^l \right\|_{\infty,N} \leq \left\| \underline{\mathbf{v}} - \underline{\mathbf{w}} \right\|_\infty .$$

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 $\|\cdot\|_{\infty,N}$ is applied to N-dimensional vectors (and not to $(M \times N)$-dimensional vectors like $\|\cdot\|_\infty$). This completes the analysis of the Deferred Correction applied to the context of the systems of ordinary differential equations. In the following we will see how to apply it to some numerical methods for hyperbolic systems of balance laws.

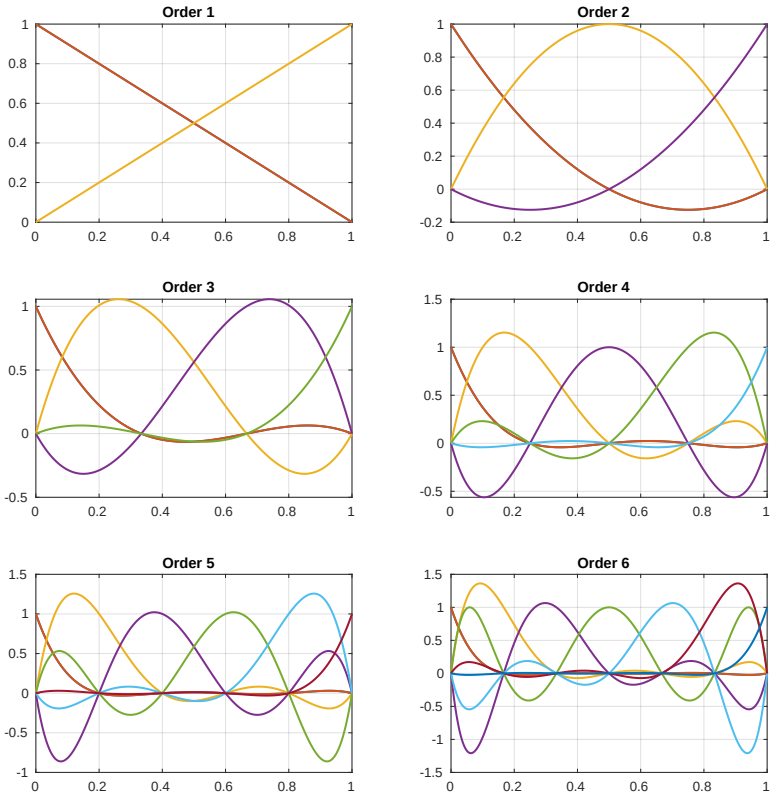


Figure 3.2: Lagrange polynomials in the reference interval $[0, 1]$.

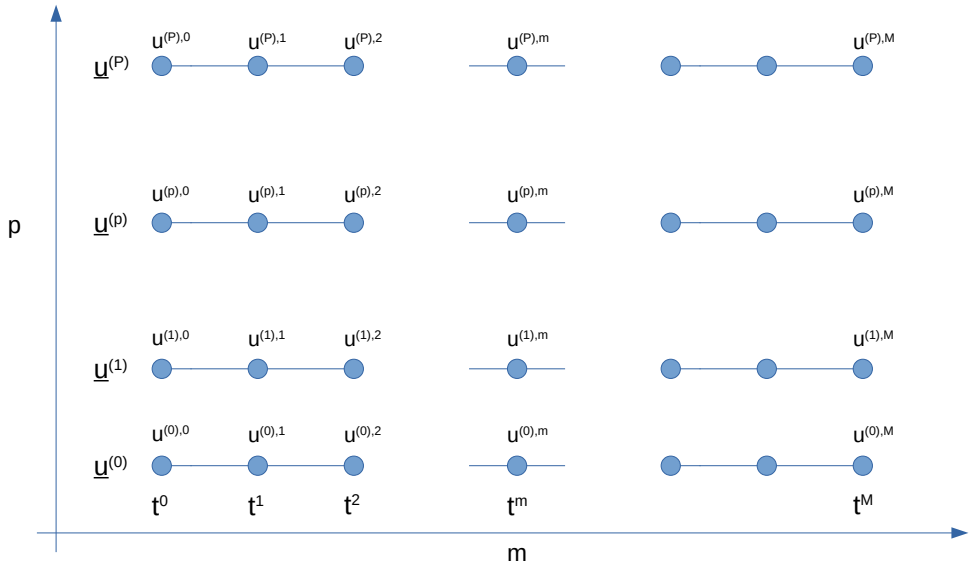


Figure 3.3: Grid of the iterations.

Chapter 4

An introduction to balance laws

In order to apply the Deferred Correction to some numerical methods for solving hyperbolic systems of balance laws, let's introduce the analytical problem. We would like to solve the general system

$$\frac{\partial}{\partial t}\mathbf{u}(\mathbf{x},t) + div_{\mathbf{x}}\mathbf{F}(\mathbf{u}(\mathbf{x},t)) = \mathbf{S}(\mathbf{x},t,\mathbf{u}(\mathbf{x},t)) \quad (\mathbf{x},t) \in \Omega \times \mathbb{R}_0^+ \tag{4.1}$$

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

The system 4.1 is a vectorial partial differential equation i.e. a set of coupled scalar partial differential equations and we have that

- $\Omega \subseteq \mathbb{R}^D$ is the spatial domain which is assumed to be a connected open set with dimension $D \in \mathbb{N}$ and its boundary $\partial\Omega$ is assumed to be smooth-enough;
- $\mathbf{u} : \overline{\Omega} \times \mathbb{R}_0^+ \to \mathbb{R}^N$ is the unknown solution with $N \in \mathbb{N}$ number of scalar equations of the system;
- $\mathbf{F} : \mathbb{R}^N \to \mathbb{R}^N \times \mathbb{R}^D$ is the flux;
- $\mathbf{S} : \overline{\Omega} \times \mathbb{R}_0^+ \times \mathbb{R}^N \to \mathbb{R}^N$ is the source (or sink) term;
- $\mathbf{u}_0 : \Omega \to \mathbb{R}^N$ is the known initial condition.

By $div_{\mathbf{x}}\mathbf{F}(\mathbf{u}(\mathbf{x},t))$ we mean the divergence operator in the only spatial coordinates applied to the flux i.e.

$$div_{\mathbf{x}}\mathbf{F}(\mathbf{u}(\mathbf{x},t)) = \sum_{d=1}^{D} \frac{\partial}{\partial x^d}\mathbf{F}_d(\mathbf{u}(\mathbf{x},t))$$

where $\mathbf{F}_d \in \mathbb{R}^N$ $d = 1, 2, \ldots, D$ is the d component of the flux $\mathbf{F} = (\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_D)$. For the sake of compactness in the following we will often refer to $\mathbf{u}(\mathbf{x},t)$, $\mathbf{F}(\mathbf{u}(\mathbf{x},t))$ and $\mathbf{S}(\mathbf{x},t,\mathbf{u}(\mathbf{x},t))$ simply by $\mathbf{u}$, $\mathbf{F}$ and $\mathbf{S}$.

Observation 21. *To be more precise we also say that in order for the system 4.1 to be hyperbolic we need that any real linear combination of the Jacobians of the components of the flux calculated in any admissible*[1] $\mathbf{u}$ *must be real diagonalizable i.e. the matrix*

$$\sum_{d=1}^{D} \omega_d \frac{\partial \mathbf{F}_d}{\partial \mathbf{u}}(\mathbf{u})$$

must have D real eigenvalues and D real eigenvectors for any admissible $\mathbf{u}$ *and* $\forall \omega = (\omega_1, \ldots, \omega_D)^T \in \mathbb{R}^D$.

4.1 Derivation of balance laws

We will now give a practical meaning to our analytical problem 4.1 by presenting its general derivation in the context of a generic real application. There is a huge number of phenomena in physics, chemistry, biology, engineering and social sciences that can be modeled through a vectorial equation like the one in 4.1. Basically whenever we want to express the conservation of some quantities (the components of the vector of the conserved quantities $\mathbf{u}$) over a certain volume Ω (either physical or not) in a Eulerian formulation we end up with a hyperbolic system of balance laws. We start by an integral balance at a general time $t \in \mathbb{R}_0^+$ for a general control volume $V \subseteq \Omega$ with respect to the quantities of interest collected in the vector $\mathbf{u}$. The situation is represented in figure 6.3. The set V is an arbitrary connected open subset of Ω with smooth-enough boundary. The balance reads

$$\int_V \frac{\partial}{\partial t}\mathbf{u} d\mathbf{x} + \int_{\partial V} \mathbf{F} \cdot \nu d\sigma(\mathbf{x}) = \int_V \mathbf{S} d\mathbf{x} \quad \forall V \subseteq \Omega, \forall t \in \mathbb{R}_0^+ \tag{4.2}$$

where ν is the outward-pointing normal vector to the surface ∂V. The equation 4.2 is known as the "integral" or "global" formulation of a system of balance laws. The first term represents the rate of change of the amount of conserved quantities $\mathbf{u}$ in the control volume V. The second term represents the surface integral of the normal flux across ∂V i.e. the amount of $\mathbf{u}$ exiting from the control volume per unit of time. The last term represents the amount of conserved quantities $\mathbf{u}$ generated by the source in the control volume per unit of time. If we assume $\mathbf{F}$ to be regular in $\overline{V}$, we can apply the Gauss divergence theorem and get

$$\int_V \frac{\partial}{\partial t}\mathbf{u} d\mathbf{x} + \int_V div_{\mathbf{x}} \mathbf{F} d\mathbf{x} = \int_V \mathbf{S} d\mathbf{x}.$$

[1] If we consider a system of balance laws modeling some real phenomenon, it can happen that not all the states of the space $\mathbb{R}^N$ are admissible for a solution $\mathbf{u}$ under a physical point of view. For example if we consider the Euler equation we have that the density and the pressure have to be positive. This results in a restriction on the space of the possible values that can be assumed from the vector $\mathbf{u}$ if we set initial and boundary conditions consistent with the physics of the problem associated to the system of balance laws.

Since this balance holds for any time $t \in \mathbb{R}_0^+$ and for any control volume $V \subseteq \Omega$, under the assumption of regularity for the integrand functions, from the arbitrariness of V we get the vectorial partial differential equation 4.1 that we recall for clarity

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

which is referred as the "differential" or "local" formulation of the system of balance laws.

Observation 22. *The control volume V, as well as the spatial domain Ω, can either be fixed or depend on time. Clearly in the first case the normal ν just depends on space while in the second case it also depends on time. Let's consider for example a deformable body with its shape varying in time, in this case the spatial domain changes in time. Anyway we will just consider spatial domains fixed in time.*

Observation 23. *So far the spatial domain Ω can also be unbounded. Clearly when we will deal with the problem under a numerical point of view we will be interested in bounded spatial domains and we will have to set a final time T.*

Several boundary conditions can be imposed. In the applications they are a representation of the physical conditions that we want to reproduce. In these cases we can derive them from the physical model. Due to the multiplicity of the existing types of boundary conditions we do not treat their derivation in detail.

4.2 Weak solutions

We will now make a small description of the solutions that we expect for a system like 4.1. A "classical" or "strong" solution to 4.1 is a smooth-enough vectorial function satisfying it pointwise (and obviously satisfying also the initial and the bounndary conditions). It is well known that the notion of classical solution is not "sufficient": for nonlinear fluxes we can get discontinuities in the solution even if we start from smooth initial conditions. In the general case there is no hope to find a strong solution. Since the mentioned discontinuities show up also in the phenomena that we want to model[2] this is not a negligible problem and we need to solve it. Let's notice also that the discontinuities in the solution determine a violation of the hypoteses of regularity under which we passed from the integral formulation to the differential one. To deal with this problem we introduce the weak solutions. We look at our equation 4.1 under a distributional point of view.

In practice we multiply our equation 4.1 by a scalar smooth arbitrary test function with compact support and integrate in space and time

$$\int_{\Omega\times\mathbb{R}_0^+} \left(\frac{\partial}{\partial t}\mathbf{u} + div_{\mathbf{x}}\mathbf{F}\right)\varphi(\mathbf{x}, t)d\mathbf{x}dt = \int_{\Omega\times\mathbb{R}_0^+} \mathbf{S}\varphi(\mathbf{x}, t)d\mathbf{x}dt \quad \forall\varphi \in C_o^1(\overline{\Omega} \times \mathbb{R}_0^+).$$

[2]Let's consider for example the shock waves generated by an aircraft moving at supersonic speed or discontinuities in the properties of non-homogeneous media.

Under hypoteses of regularity of the solution and the flux we can use the divergence theorem and write

$$\int_\Omega \left([\mathbf{u}\varphi(\mathbf{x},t)]_{t=0}^{t=+\infty} - \int_{\mathbb{R}_0^+} \mathbf{u}\frac{\partial}{\partial t}\varphi(\mathbf{x},t)dt \right) d\mathbf{x} + \int_{\mathbb{R}_0^+} \left(\int_{\partial\Omega} \varphi(\mathbf{x},t)\mathbf{F}\cdot\nu d\sigma(\mathbf{x}) - \int_\Omega \mathbf{F}\cdot\nabla_{\mathbf{x}}\varphi(\mathbf{x},t)d\mathbf{x} \right) dt =$$

$$= \int_{\Omega\times\mathbb{R}_0^+} \mathbf{S}\varphi(\mathbf{x},t)d\mathbf{x}dt \quad \forall\varphi \in C_o^1(\overline{\Omega}\times\mathbb{R}_0^+)$$

where $\nabla_{\mathbf{x}}$ is the gradient in the only spatial coordinates and ν is the outward-pointing normal vector to the surface $\partial\Omega$. Thus due to the compact support of φ and reminding the initial condition $\mathbf{u}(\mathbf{x},0) = \mathbf{u}_0(\mathbf{x})$ we have

$$-\int_\Omega \mathbf{u}_0(\mathbf{x})\varphi(\mathbf{x},0)d\mathbf{x} - \int_{\Omega\times\mathbb{R}_0^+} \left(\mathbf{u}\frac{\partial}{\partial t}\varphi(\mathbf{x},t) + \mathbf{F}\cdot\nabla_{\mathbf{x}}\varphi(\mathbf{x},t) \right) d\mathbf{x}dt +$$

$$+ \int_{\partial\Omega\times\mathbb{R}_0^+} (\varphi(\mathbf{x},t)\mathbf{F}\cdot\nu)\, d\sigma(\mathbf{x})dt = \int_{\Omega\times\mathbb{R}_0^+} \mathbf{S}\varphi(\mathbf{x},t)d\mathbf{x}dt \quad \forall\varphi \in C_o^1(\overline{\Omega}\times\mathbb{R}_0^+). \tag{4.3}$$

Equation 4.3 is referred as the weak formulation of the system of balance laws 4.1.

Observation 24. *In reality the integral over the surface $\partial\Omega\times\mathbb{R}_0^+$ (and so the boundary conditions) would require more rigorous specifications under an analytic point of view. More in general we admit that a more formal position of the problem is possible in the context of the distributions i.e. linear and continuous (w.r.t. a certain notion of convergence) operators over a space of test functions but it is out of the purpose of this document.*

A weak solution is a function $\mathbf{u} : \overline{\Omega}\times\mathbb{R}_0^+ \to \mathbb{R}^N$ satisfying the integral relation 4.3 for any arbitrary test function $\varphi \in C_o^1(\overline{\Omega}\times\mathbb{R}_0^+)$. To a certain extent we are going back to an integral formulation like the one presented in equation 4.2 but slightly different. We actually got the differential formulation 4.1 from the integral one 4.2 under some regularity assumptions which, as we said, are not often reliable. A return to an integral formulation is abundantly justified. Moreover the weak formulation require less regularity from a possible solution. If we look at 4.3, all the derivatives are now on the smooth test function, even a discontinuous function can be a weak solution.

Observation 25. *We can simply verify that a strong solution is also a weak solution. Clearly the viceversa is not true since the weak solutions can have discontinuities. Moreover it is also easy to see that if a weak solution is smooth in a neighborhood of an arbitrary point $(\mathbf{x},t)$ of the space-time domain then it satisfies in the strong sense the initial equation in that point.*

Observation 26. *If we take $\Omega = \mathbb{R}^D$, due to the compact support of φ, we have that 4.3 reduces to*

$$-\int_{\mathbb{R}^D} \mathbf{u}_0(\mathbf{x})\varphi(\mathbf{x},0)d\mathbf{x} - \int_{\mathbb{R}^D\times\mathbb{R}_0^+} \left(\mathbf{u}\frac{\partial}{\partial t}\varphi(\mathbf{x},t) + \mathbf{F}\cdot\nabla_{\mathbf{x}}\varphi(\mathbf{x},t) \right) d\mathbf{x}dt =$$

$$= \int_{\mathbb{R}^D \times \mathbb{R}_0^+} \mathbf{S}\varphi(\mathbf{x}, t) d\mathbf{x} dt \quad \forall \varphi \in C_o^1(\mathbb{R}^D \times \mathbb{R}_0^+).$$

Many other things can be said about weak solutions: it can be shown that a shock (a discontinuity of first kind) in a weak solution must fulfil the Rankine-Hugoniot condition which represents a relation between the "jump" and the "speed"; the weak solutions are in general not unique so we need a criterion to select the ones which have a physical meaning and this is often given by some inequalities, the weak solutions satisfying these inequalities are said "entropy solutions". A deeper analysis is out of the purpose of this document. What we have said so far can seem to be very distant from our initial goal of showing how to apply the Deferred Correction to the numerical methods for systems of balance laws but in the author's opinion it was fundamental in order to give a clear idea of the problem that we are trying to solve. Moreover almost all the numerical approaches are designed to detect approximations of weak (eventually entropic) solutions.

We can now enter into the details of the numerical methods for hyperbolic systems of balance laws.

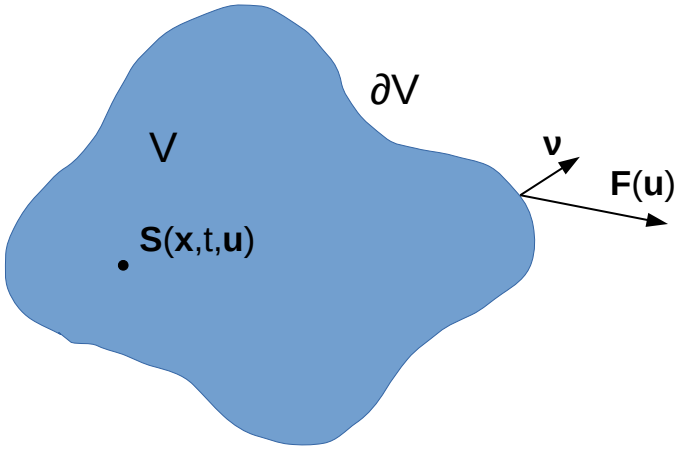


Figure 4.1: Representation of the conservation in the general control volume V.

Chapter 5

The semidiscrete formulation and the role of the mass matrix

Before presenting the Residual Distribution, or simply RD, and the Galerkin Finite Elements methods[1], or simply Galerkin FEMs, and show how to apply the Deferred Correction in the context of these ones, we introduce from a general point of view the semidiscrete formulation which is a general strategy shared by several numerical methods for hyperbolic systems of balance laws[2] (among which Residual Distribution and Galerkin Finite Elements) to deal separately with the space and time discretizations. It leads to a system of ODEs commonly involving a mass matrix. We will focus on the problem of the inversion of this matrix. We remark that the application of the Deferred Correction to the Residual Distribution and polynomial Galerkin Finite Elements methods allows to avoid the inversion of the mass matrix leading to explicit schemes.

We would like to solve numerically a hyperbolic system of balance laws in the general form 4.1 that we recall for clarity

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

with some initial and boundary conditions. As anticipated, from a numerical point of view we are interested in bounded spatial domains Ω and we have to choose a final time $T \in \mathbb{R}^+$. Thus we would like to find $\mathbf{u}_h$ approximation in $\overline{\Omega} \times [0, T]$ of $\mathbf{u}$ which is a weak (eventually entropic) solution to our problem i.e. satisfying the integral relation 4.3 with $\Omega \subseteq \mathbb{R}^D$ bounded connected open set with a smooth-enough boundary (and eventually some entropy inequalities).

Observation 27. *Often the reference spatial domain for the numerical method is just an approximation of the analytical spatial domain Ω in which we would like to solve the initial equation. This reference domain is usually referred as "discretized". In this work*

[1] We will just present the methods without treating the problem of the convergence of the detected approximated solutions to the desired analytical weak ones.

[2] In reality the semidiscrete formulation also applies to some numerical methods for other types of equations but we will just consider hyperbolic systems of balance laws.

we assume for simplicity that the discretized domain coincides with the analytical one. Thus we use indifferently Ω for the "analytical" and "numerical" reference domains and so we will do when we will enter into the details of the Residual Distribution and the Galerkin Finite Elements methods.

A common approach to solve numerically a system of balance laws consists in "splitting" the resolution of the equation in space and the resolution of the equation in time. We assume at any time $t \in [0, T]$ the approximated solution $\mathbf{u}_h$ in a finite dimensional space $(V_h)^N$ where $V_h = span\{\varphi_i\}_{i=1,2,\dots,I}$ and the (known) functions $\varphi_i : \overline{\Omega} \to \mathbb{R}$ are a basis of V_h and depend just on space.

Observation 28. *The functions φ_i can be either continuous or discontinuous and are usually but not mandatorily piecewise polynomial functions. In the case of piecewise polynomial approximations, we deal usually but not mandatorily with "nodal" bases (like for example the Lagrange and the Bernstein bases). These ones are such that each basis function φ_i corresponds to a node $\mathbf{x}_i$ located somewhere in the closure of the discretized spatial domain. In these cases we have that the value of the basis function φ_i in its associated node $\mathbf{x}_i$ is not 0 i.e. $\varphi_i(\mathbf{x}_i) \neq 0$. The nodes $\mathbf{x}_i$ are often referred as the degrees of freedom, DoFs, of the approximated solution. If we have continuous piecewise polynomial approximations to different basis functions are associated different nodes, while in the context of discontinuous piecewise polynomial approximations to each node are in general associated different basis functions. We will at first introduce the methods from a general point of view but we say in advance that we will apply the Deferred Correction just in the case of piecewise polynomial nodal bases so we will then focus on these ones.*[3].

Coming back to our approximation, in other words we require every component of the approximated solution at any time t to be in the space of functions V_h. In practice we are assuming

$$\mathbf{u}_h(\mathbf{x}, t) = \sum_{i=1}^{I} \mathbf{c}_i(t)\varphi_i(\mathbf{x}) \quad \forall (\mathbf{x}, t) \in \overline{\Omega} \times [0, T] \tag{5.1}$$

with $\mathbf{c}_i(t) \in \mathbb{R}^N \quad \forall i = 1, 2, \dots, I \quad \forall t \in [0, T]$.

Observation 29. *The basis functions φ_i are fixed and known, the only unknowns are the coefficients $\mathbf{c}_i(t)$. Looking for $\mathbf{u}_h$ is equivalent to look for $\mathbf{c}_i(t) \quad i = 1, 2, \dots, I \quad t \in [0, T]$.*

The problem is now to evaluate these coefficients in $[0, T]$. Thus, according to some mathematical and physical criteria depending on the chosen method, some relations based on the initial equation 4.1 are derived. From the combination of these relations

[3]This observation is actually referred to the Galerkin Finite Elements methods in which the choice of nonpolynomial bases is not so unusual while in the Residual Distribution framework we always deal with continuous piecewise polynomial approximations.

and the discretization 5.1 we usually get a system of ordinary differential equations involving a mass matrix $\mathcal{M}$ which reads

$$\mathcal{M}\frac{d}{dt}\mathbf{c}(t) = \mathbf{H}(t, \mathbf{c}(t)) \tag{5.2}$$

where the unknown vector $\mathbf{c}(t)$ has $I \times N$ dimensions and contains all the N-dimensional unkonwn vectors $\mathbf{c}_i(t)$ as components

$$\mathbf{c}(t) = \begin{pmatrix} \mathbf{c}_1(t) \\ \vdots \\ \mathbf{c}_i(t) \\ \vdots \\ \mathbf{c}_I(t) \end{pmatrix} \in \mathbb{R}^{I \times N}.$$

We have obviously $\mathcal{M} \in \mathbb{R}^{(I\times N)\times(I\times N)}$ and the vector $\mathbf{H}(t, \mathbf{c}(t)) \in \mathbb{R}^{(I\times N)}$ $\forall t \in [0, T]$. The system of ODEs 5.2 is usually referred as "semidiscrete formulation" of the particular method that we are using and we have to solve it in time. The initial condition $\mathbf{c}(0) = \mathbf{c}_0$ is given by the analytical initial condition on the hyperbolic system of balance laws 4.1 that we want to solve (projected into the finite dimensional space $(V_h)^N$).

Observation 30. *The Galerkin Finite Elements and the Residual Distribution methods are based on this approach, so in both cases we get a system of coupled ODEs like the one in equation 5.2. In the Discontinuous Galerkin case the mass matrix is not dependent on time nor on* $\mathbf{c}(t)$*, moreover it is defined positive (thus nonsingular) and block diagonal. Usually in this case the mass matrix is inverted and the resulting system of ODEs*

$$\frac{d}{dt}\mathbf{c}(t) = \mathcal{M}^{-1}\mathbf{H}(t, \mathbf{c}(t))$$

is solved with a classical ODE solver, typically a Runge-Kutta scheme[4]*. In the general case the matrix* $\mathcal{M}$ *depends on the time* t *and on the vector* $\mathbf{c}(t)$ *and we don't even know whether it is invertible or not. This is the case for example of some Residual Distribution methods. In such occurrences trying to solve the system as in the Discontinuous Galerkin case would lead to very big problems*[5] *thus avoiding the inversion of the mass matrix is crucial. Some other Residual Distribution methods lead to a semidiscrete formulation very close to the one of the stabilized Continuous Galerkin methods for piecewise polynomial nodal bases with a mass matrix not dependent on time nor on the solution, defined*

[4]To be more specific we also say that a huge variety of ODEs solvers exists and depending on the features that we are interested in we can choose particular methods rather than others. For example if we are interested in controlling the total variation of the solution we can choose a strong stability preserving Runge-Kutta method (SSP-RK). If we are interested in the preservation of the non-negativity of certain variables we can choose a positivity preserving scheme. We do not enter into the details of such methods.

[5]As we said, we have no information about the invertibility of the matrix. Moreover since it depends on t and on $\mathbf{c}(t)$, a classical approach would be very likely to require to recompute it and invert it (assuming that this is possible) at each time step.

positive (thus nonsingular) but sparse. The inversion of the matrix is not cheap by a computational point of view because it is sparse and we cannot rely on the efficient algorithms designed for block diagonal matrices contrarily to the Discontinuous Galerkin case. The reader could argue that this operation has to be done done just once. This is indeed true but if we opt for a standard method requiring the inversion of the mass matrix, at each time step we will have to multiply this huge inverse for some evaluations of the vector $\mathbf{H}$ *(or its approximations). This results in a very big computational cost. Here the Deferred Correction comes to aid. We will see how to apply it in the context of these methods treating at the same time these last mentioned Residual Distribution methods and the stabilized Continuous Galerkin ones with piecewise polynomial nodal bases. Then we will see how, in a similar fashion, the same technique can be applied to the Discontinuous Galerkin case again considering piecewise polynomial nodal bases.*

Observation 31. *The Finite Volume methods are characterized by a different approach. The coefficients represent the averages of the approximated solution in some regions of the (discretized) space domain but also in this case we arrive to a semidiscrete formulation like equation 5.2. In this case the mass matrix is diagonal, nonsingular and its entries are the measures of the mentioned regions. Thus in this case the inversion of the mass matrix is straightforward.*

Observation 32. *Generally a spatial discretization made with polynomial functions* φ_i *of order* R *and an ODE solver* $(R+1)$*-order accurate lead to a globally* $(R+1)$*-order accurate scheme. This means that the coefficients* $\tilde{\mathbf{c}}(t^*)$ *that characterize the approximated solution* $\mathbf{u}_h$ *at the time* t^* *determined through the used method are* $(R+1)$*-order accurate approximations of the coefficients* $\mathbf{c}(t^*)$ *that we obtain projecting the exact solution at the time* t^* *over* $(V_h)^N$*. But let's underline that a spatial discretization made with polynomial functions of order* R *and an ODE solver* $(R+1)$*-order accurate are just necessary but not sufficient ingredients for* $(R+1)$*-order accuracy. If the equations of the system 5.2 are not carefully derived, the accuracy is wasted and we get unexpected low orders of convergence.*

Let's now present more in detail the Residual Distribution and the Continous and Discontinuous Galerkin Finite Elements methods and see how to apply the Deferred Correction to get explicit arbitrary high order accurate schemes for hyperbolic systems of balance laws.

Chapter 6

The Residual Distribution and the stabilized Continuous Galerkin Finite Elements methods

We start by presenting the Residual Distribution method, then we introduce the stabilized Continuous Galerkin FEM and in the end we show the connections between these approaches and we write the reference semidiscrete formulation for the application of the Deferred Correction to get explicit schemes avoiding the inversion of the mass matrix.

6.1 Residual Distribution

We refer again to our initial problem. We would like to find an approximation $\mathbf{u}_h$ in $\overline{\Omega} \times [0, T]$ of a weak (eventually entropic) solution of an hyperbolic system of balance laws whose general form is given by equation 4.1 which is recalled here

$$\frac{\partial}{\partial t}\mathbf{u} + div_{\mathbf{x}}\mathbf{F} = \mathbf{S} \quad \forall(\mathbf{x}, t) \in \Omega \times \mathbb{R}_0^+.$$

Again we remark that now Ω is a bounded domain with smooth-enough boundary $\partial\Omega$. Some suitable initial and boundary conditions are assumed. We consider a tessellation $\mathcal{T}_h$ of the closure of the spatial domain i.e. a finite family of D-dimensional nonoverlapping convex polytopal subsets K of $\overline{\Omega}$ which cover it exactly[1] with characteristic length h, i.e. such that

- $\mathring{K} \neq \emptyset \quad \forall K \in \mathcal{T}_h$
- $\mathring{K_i} \cap \mathring{K_j} = \emptyset \quad \forall K_i, K_j \in \mathcal{T}_h \quad s.t. \quad K_i \neq K_j$;

[1] Again we remind the assumption that the analytical and the discretized spatial domain coincide. The elements of the tessellation cover $\overline{\Omega}$ exactly without any approximation.

- $\cup_{K\in\mathcal{T}_h} K = \overline{\Omega}$.
- $\sup_{K\in\mathcal{T}_h} diam(K) = h$ where $diam(K) = \sup_{\mathbf{x},\mathbf{y}\in K} \|\mathbf{x}-\mathbf{y}\|_2$

Observation 33. *The shape of the sets of the tessellation cannot be chosen arbitrarily, there are some regularity requirements that must be satisfied but we do not treat them in detail.*

The tessellation $\mathcal{T}_h$ is often called "triangulation" and its elements are ususally referred as "cells". Usually but not mandatorily the polytopes of the tessellation are simplices. In figure 6.1 are represented the simplices in one, two and three dimensions.

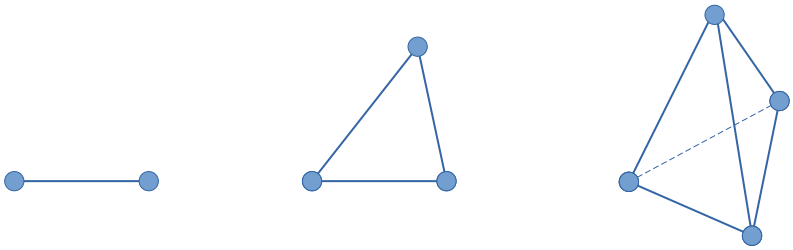
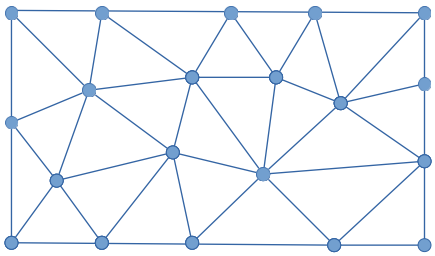


Figure 6.1: Simplices in one, two and three dimensions.

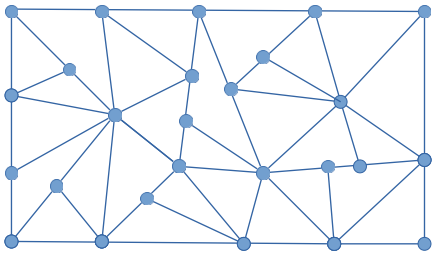
Usually but not mandatorily the tessellation is "conformal". Roughly speaking it means that each $D-1$-dimensional polytopal face f belongs at most to two polytopes. For example, for $D = 3$ this is equivalent to say that two different elements of the tessellation can share at most one vertex or a whole edge or a whole face while for $D = 2$ they can share at most one vertex or a whole edge (i.e. a 1-dimensional face). To be clearer have a look at the picture 6.2 in which a conformal and a non-conformal bidimensional tessellations are shown.

We will refer to a conformal tessellation made by simplices but our description will be in general valid also for more general tessellations. To get the semidiscrete formulation we assume that the approximated solution $\mathbf{u}_h$ is at any time $t \in [0,T]$ in the space $(V_h)^N$ where V_h is a space of continuous piecewise-polynomial functions defined as

$$V_h = \{g \in C^0(\overline{\Omega}) \quad s.t. \quad g|_K \in \mathbb{P}_M \quad \forall K \in \mathcal{T}_h\}.$$



(a) Conformal tessellation.



(b) Non conformal tessellation.

Figure 6.2: Example of tessellations of a bidimensional rectangular domain.

To this end we choose a basis $\{\varphi_i\}_{i=1,2,\dots,I}$ of V_h, for example the Lagrange polynomials or the Bernstein polynomials, and we assume the discretization 5.1 that is recalled here

$$\mathbf{u}_h(\mathbf{x},t) = \sum_{i=1}^{I} \mathbf{c}_i(t)\varphi_i(\mathbf{x}) \quad \forall(\mathbf{x},t) \in \overline{\Omega} \times [0,T]$$

with $\mathbf{c}_i(t)$ N-dimensional coefficients dedpending on time. The general basis function φ_i is associated to a node $\mathbf{x}_i \in \overline{\Omega}$ and is such that $\varphi_i(\mathbf{x}_i) \neq 0$. These I nodes associated to the I basis functions are the degrees of freedom of the solution and belong in general to one or more elements of the tessellation $\mathcal{T}_h$. Each basis function φ_i has its support contained in the union of the elements containing the node to which it is associated. In other words if φ_i is associated to the node $\mathbf{x}_i \in \overline{\Omega}$ and we denote with K_i the set of the elements of the tessellation containing the node $\mathbf{x}_i$ i.e.

$$K_i = \{K \in \mathcal{T}_h \quad s.t. \quad \mathbf{x}_i \in K\}$$

then $\varphi_i \in C^0_o(\cup_{K \in K_i} K)$. The situation is represented in figure 6.3: three nodes are put in evidence and coloured respectively in green, yellow and red as well as the elements of the associated K_i i.e. the elements containing them which are coloured with the same colours but in a lighter shade.

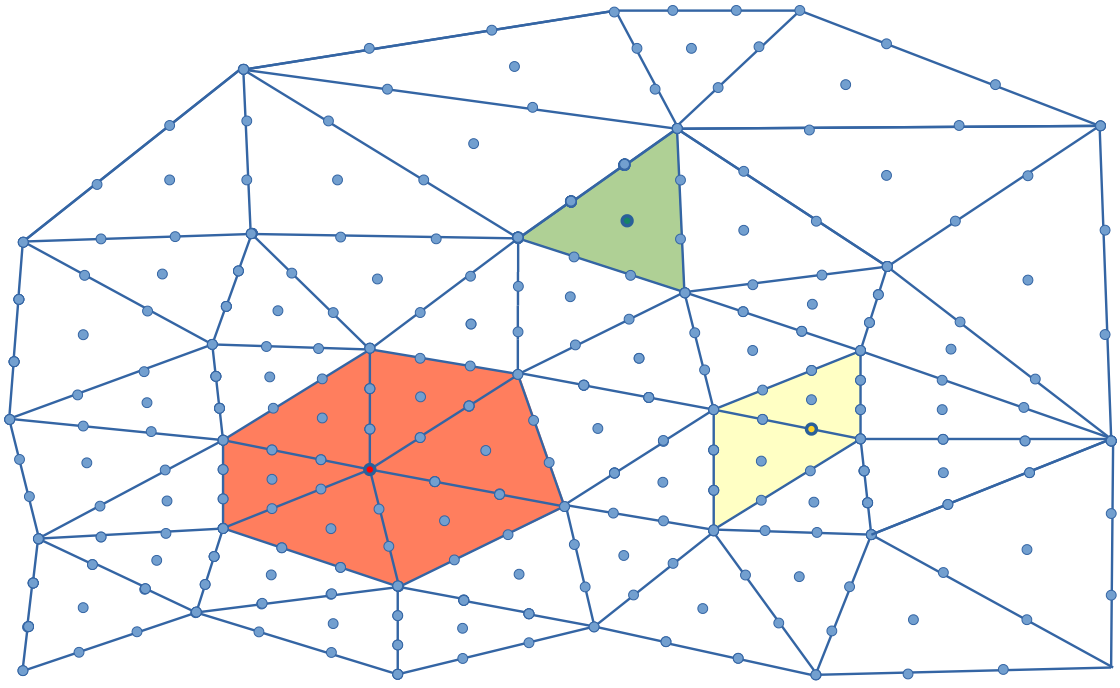


Figure 6.3: Some degrees of freedom and the associated K_i.

For example, the basis function associated to the red degree of freedom has its support contained in the union of the six red elements of the tessellation which contain it. The

basis function associated to the yellow node is non-zero just in the union of the two yellow elements containing it. And obviosly the basis function associated to the green degree of freedom has support contained in the only (green) element to which the green node belongs. Moreover in the particular case of the Residual Distribution approach we always normalize the basis functions s.t. their sum is identically equal to 1 over $\overline{\Omega}$ i.e.

$$\sum_{i=1}^{I} \varphi_i(\mathbf{x}) \equiv 1 \quad \forall \mathbf{x} \in \overline{\Omega}. \tag{6.1}$$

Observation 34. *This is usually but not mandatorily done also for other methods.*

Observation 35. *The Lagrangian basis functions are characterized by the fact that their value is* 1 *in the node to which they are associated and* 0 *in all the other nodes i.e.*

$$\varphi_i(\mathbf{x}_j) = \delta_{ij} \quad \forall i, j = 1, 2, \cdots, I$$

where δ_{ij} is the Kronecker delta. Thus in this case the coefficients $\mathbf{c}_j(t)$ of the discretization 5.1 represent directly the value of the components of the approximated solution $\mathbf{u}_h$ in the node $\mathbf{x}_j$ at the time t

$$\mathbf{u}_h(\mathbf{x}_j, t) = \sum_{i=1}^{I} \mathbf{c}_i(t)\varphi_i(\mathbf{x}_j) = \sum_{i=1}^{I} \mathbf{c}_i(t)\delta_{ij} = \mathbf{c}_j(t) \quad \forall t \in [0, T].$$

This is in general not true for other choices of the basis, for example for the Bernstein polynomials.

Now we have all the elements that we need to go into the details of the Residual Distribution approach.

Basically the Residual Distribution can be summarised into three steps:

i) **Definition of the element residuals**
For each element K of the tessellation $\mathcal{T}_h$ we define the element residual $\mathbf{\Phi}^K(t, \mathbf{u}_h)$ which is what we get if we make the integral in space of the equation 4.1 over the element K and apply the divergence theorem

$$\forall K \in \mathcal{T}_h \quad \mathbf{\Phi}^K(t, \mathbf{u}_h) = \int_K \frac{\partial}{\partial t}\mathbf{u}_h(\mathbf{x}, t)d\mathbf{x} + \int_{\partial K} \mathbf{F}(\mathbf{u}_h(\mathbf{x}, t)) \cdot \nu d\sigma(\mathbf{x}) + \\ - \int_K \mathbf{S}(\mathbf{x}, t, \mathbf{u}_h(\mathbf{x}, t))d\mathbf{x} \quad t \in [0, T]; \tag{6.2}$$

ii) **Definition of the node residuals**
For each element K we consider the degrees of freedom belonging to it, $\mathbf{x}_i \in K$, and define the node residuals $\mathbf{\Phi}_i^K(t, \mathbf{u}_h)$ such that they satisfy a specific conservation relation

$$\forall K \in \mathcal{T}_h \quad \forall \mathbf{x}_i \in K \quad \mathbf{\Phi}_i^K(t, \mathbf{u}_h) \quad t \in [0, T] \quad s.t. \\ \sum_{\mathbf{x}_i \in K} \mathbf{\Phi}_i^K(t, \mathbf{u}_h) = \mathbf{\Phi}^K(t, \mathbf{u}_h) \quad \forall K \in \mathcal{T}_h \quad \forall t \in [0, T]; \tag{6.3}$$

iii) **Imposition of the balance at the nodes**
For each degree of freedom $\mathbf{x}_i$ we impose an equilibrium between all the node residuals $\mathbf{\Phi}_i^K(t, \mathbf{u}_h)$ of the elements that contain that degree of freedom

$$\sum_{K \in K_i} \mathbf{\Phi}_i^K(t, \mathbf{u}_h) = \mathbf{0} \quad t \in [0, T] \quad \forall \imath = 1, 2, \ldots, I \tag{6.4}$$

where we recall that K_i is the set of the elements of the tessellation containing the node $\mathbf{x}_i$. The balance 6.4 represents the semidiscrete formulation of the Residual Distribution method, a system of nonlinear ODEs that we have to solve in time.

We will now give a physical interpretation of the Residual Distribution approach. Let's go back to the steps of the method and look at them under a new perspective.

i) **Balances at the elements**
We start by an initial balance at each element. We consider our initial equation 4.1 and we move the source at the left hand side so to have

$$\frac{\partial}{\partial t}\mathbf{u}(\mathbf{x}, t) + div_{\mathbf{x}}\mathbf{F}(\mathbf{u}(\mathbf{x}, t)) - \mathbf{S}(\mathbf{x}, t, \mathbf{u}(\mathbf{x}, t)) = \mathbf{0}.$$

If we integrate this quantity over each element K of the tessellation at a generic time t and we apply the divergence theorem we get the integral balance

$$\int_K \frac{\partial}{\partial t}\mathbf{u}(\mathbf{x}, t)d\mathbf{x} + \int_{\partial K} \mathbf{F}(\mathbf{u}(\mathbf{x}, t)) \cdot \nu d\sigma(\mathbf{x}) - \int_K \mathbf{S}(\mathbf{x}, t, \mathbf{u}(\mathbf{x}, t))d\mathbf{x} = \mathbf{0} \quad \forall t \in \mathbb{R}_0^+.$$

This is the meaning of the element residuals $\mathbf{\Phi}^K(t, \mathbf{u}_h)$ even if obviously to evaluate them in our numerical method we consider the approximated solution $\mathbf{u}_h(\mathbf{x}, t) = \sum_{i=1}^I \mathbf{c}_i(t)\varphi_i(\mathbf{x})$ in place of the exact solution $\mathbf{u}$.

ii) **Isolation of the contributions of the nodes to the element balance**[2]
We isolate the contribution of each degree of freedom $\mathbf{x}_i$ in the cell K to the integral balance that we have in that cell. The node residuals $\mathbf{\Phi}_i^K(t, \mathbf{u}_h)$ are nothing more than the contributions of the nodes to the balance in the cell K represented by $\mathbf{\Phi}^K(t, \mathbf{u}_h)$. This fact is expressed by the conservation relation 6.3 that the node residuals must fulfil

$$\sum_{\mathbf{x}_i \in K} \mathbf{\Phi}_i^K(t, \mathbf{u}_h) = \mathbf{\Phi}^K(t, \mathbf{u}_h) \quad \forall K \in \mathcal{T}_h \quad \forall t \in [0, T].$$

iii) **Imposition of a staggered balance at the nodes**
Once we isolated the node contributions to the element balance in each cell we

[2]This step is sometimes referred as the "splitting" of the element residual into the node residuals. Up to the author opinion it doesn't express at all the physics behind the operation, for this reason we will try to avoid this expression in the document.

impose a new "staggered equilibrium" at the nodes. The sum of the contributions that each node offers to the elements to which it belongs is 0 i.e.

$$\sum_{K \in K_i} \mathbf{\Phi}_i^K(t, \mathbf{u}_h) = \mathbf{0} \quad t \in [0,T] \quad \forall \imath = 1, 2, \ldots, I$$

This is indeed a reasonable conservation requirement: nothing is created or destroyed at the nodes. The global contribution of each node to all the balances of all elements that share it is 0.

The three steps are represented in figure 6.4.

Up to now the method is very general and also very abstract since we didn't specify how to define the node residuals $\mathbf{\Phi}_i^K(t, \mathbf{u}_h)$. The features of the particular Residual Distribution method depend on the choice of the node residuals.

Observation 36. *We have neglected the boundary conditions due to the huge variety of existing ones and the different ways to implement them. We just say that whenever they are not imposed strongly, i.e. fixing the coefficients $\mathbf{c}_i(t)$ for the nodes belonging to the boundary, they are usually taken into account by defining some boundary residuals which are added in the balances of the nodes on the boundary. This is for example what is done for the inflow-outflow boundary conditions. Anyway the structure of the Deferred Correction applied to the Residual Distribution method as well as the proofs of the properties of the operators $\mathcal{L}_\Delta^1$ and $\mathcal{L}_\Delta^2$ that we will define doesn't change much when we include the boundary conditions.*

6.1.1 Some examples of node residuals

We will now present some examples of suitable node residuals and show that they fulfil the conservation property 6.3.

- **Lax-Friedrichs**

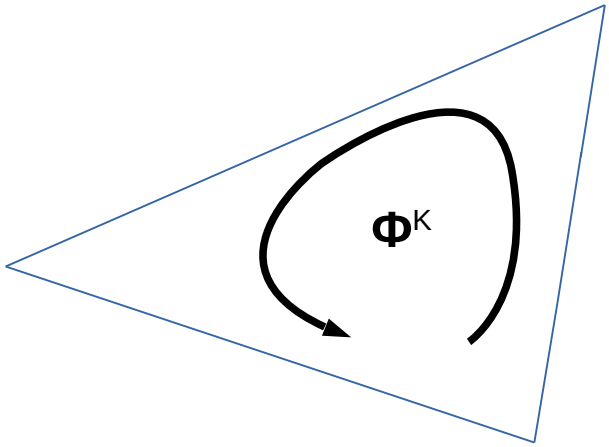
$$\mathbf{\Phi}_i^K(t, \mathbf{u}_h) = \int_K \varphi_i(\mathbf{x}) \frac{\partial}{\partial t} \mathbf{u}_h(\mathbf{x}, t) d\mathbf{x} + \int_{\partial K} \varphi_i(\mathbf{x}) \mathbf{F}(\mathbf{u}_h(\mathbf{x}, t)) \cdot \nu d\sigma(\mathbf{x}) - \int_K \mathbf{F}(\mathbf{u}_h(\mathbf{x}, t)) \cdot \nabla_{\mathbf{x}} \varphi_i(\mathbf{x}) d\mathbf{x} +$$
$$+ \alpha_K(\mathbf{u}_h) \left(\mathbf{c}_i(t) - \overline{\mathbf{c}}_K(t) \right) - \int_K \mathbf{S}(\mathbf{x}, t, \mathbf{u}_h(\mathbf{x}, t)) \varphi_i(\mathbf{x}) d\mathbf{x} \quad t \in [0,T] \tag{6.5}$$

where $\overline{\mathbf{c}}_K(t)$ is the average of the coefficients of the approximated solution associated to the nodes belonging to the cell K i.e.

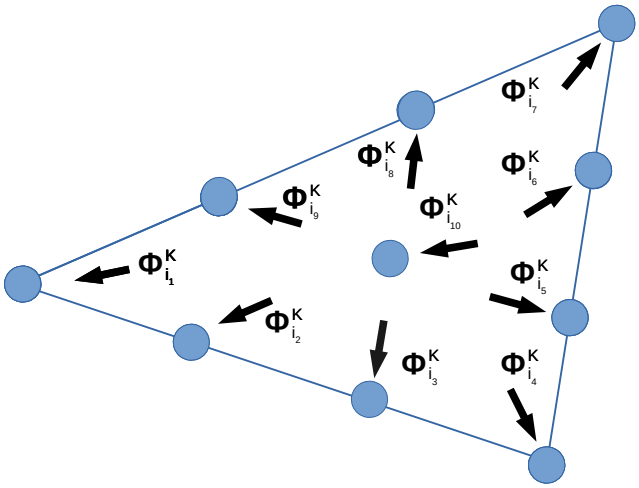
$$\overline{\mathbf{c}}_K(t) = \frac{1}{\#\{\mathbf{x}_i \in K\}} \sum_{\mathbf{x}_i \in K} \mathbf{c}_i(t)$$

and $\alpha_K(\mathbf{u}_h)$ is defined through the Jacobian matrix of the flux in the following way

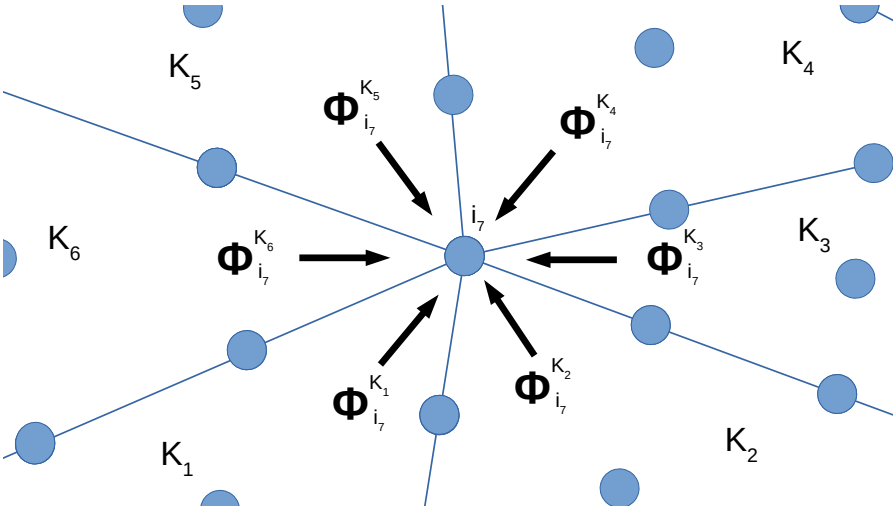
$$\alpha_K(\mathbf{u}_h) = \sup_{\substack{\mathbf{x} \in f \\ f \subset \partial K}} \rho \left(\frac{\partial \mathbf{F}}{\partial \mathbf{u}}(\mathbf{u}_h) \nu \right)$$



(a) Balance at the element K.



(b) Isolation of the contributions of the nodes to the element balance.



(c) Staggered balance at the node i_7.

Figure 6.4: The three steps of the Residual Distribution method.

where $\rho\left(\frac{\partial \mathbf{F}}{\partial \mathbf{u}}(\mathbf{u}_h)\nu\right)$ is the spectral radius of the matrix $\frac{\partial \mathbf{F}}{\partial \mathbf{u}}(\mathbf{u}_h)\nu$ i.e. the supremum of the absolute values of its eigenvalues and f is a generic $D-1$-dimensional face of the generic element K of the tessellation;

- **SUPG**

$$\begin{aligned}\mathbf{\Phi}_i^K(t,\mathbf{u}_h) = \int_K \varphi_i(\mathbf{x})\frac{\partial}{\partial t}\mathbf{u}_h(\mathbf{x},t)d\mathbf{x} + \int_{\partial K}\varphi_i(\mathbf{x})\mathbf{F}(\mathbf{u}_h(\mathbf{x},t))\cdot\nu d\sigma(\mathbf{x}) - \int_K \mathbf{F}(\mathbf{u}_h(\mathbf{x},t))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x} + \\ + h_K\int_K\left[\frac{\partial \mathbf{F}}{\partial \mathbf{u}}(\mathbf{u}_h(\mathbf{x},t))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})\right]\tau\left[\frac{\partial \mathbf{F}}{\partial \mathbf{u}}(\mathbf{u}_h(\mathbf{x},t))\cdot\nabla_{\mathbf{x}}\mathbf{u}_h(\mathbf{x},t)\right]d\mathbf{x} + \\ - \int_K \mathbf{S}(\mathbf{x},t,\mathbf{u}_h(\mathbf{x},t))\varphi_i(\mathbf{x})d\mathbf{x} \quad t\in[0,T] \end{aligned} \tag{6.6}$$

where h_K is a constant depending on the cell K and $\tau \in \mathbb{R}^+$ is a positive parameter;

- **CIP**[3]

$$\begin{aligned}\mathbf{\Phi}_i^K(t,\mathbf{u}_h) = \int_K \varphi_i(\mathbf{x})\frac{\partial}{\partial t}\mathbf{u}_h(\mathbf{x},t)d\mathbf{x} + \int_{\partial K}\varphi_i(\mathbf{x})\mathbf{F}(\mathbf{u}_h(\mathbf{x},t))\cdot\nu d\sigma(\mathbf{x}) - \int_K \mathbf{F}(\mathbf{u}_h(\mathbf{x},t))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x} + \\ + \sum_{f\subset\partial K}\theta h_f^2\int_f [\![\nabla_{\mathbf{x}}\mathbf{u}_h(\mathbf{x},t)]\!]\cdot\nabla_{\mathbf{x}}\varphi_i|_K(\mathbf{x})d\sigma(\mathbf{x}) + \\ - \int_K \mathbf{S}(\mathbf{x},t,\mathbf{u}_h(\mathbf{x},t))\varphi_i(\mathbf{x})d\mathbf{x} \quad t\in[0,T] \end{aligned} \tag{6.7}$$

where f is a generic $D-1$-dimensional face of the generic element K, $\theta \in \mathbb{R}^+$ is a positive parameter, h_f is a constant depending on the face f, $[\![\cdot]\!]$ represents the jump of the argument across f i.e.

$$[\![g(\mathbf{x})]\!] = g|_K(\mathbf{x}) - g|_{K^+}(\mathbf{x}) \quad \forall x \in f$$

with K^+ the element of the tessellation sharing the face f with K and $\cdot|_K$ and $\cdot|_{K^+}$ representing the restrictions to K and K^+ thus

$$g|_K(\mathbf{x}) = \lim_{\substack{\mathbf{y}\to\mathbf{x}\\ \mathbf{y}\in K}} g(\mathbf{y})$$

[3]The name "CIP" stays for "continuous interior penalty" and is due to the fact that the Residual Distribution method that we get using these node residuals is equivalent to a stabilized Continuous Galerkin FEM with stabilization term based on the Interior Penalty Procedure introduced by Jim Douglas Jr. and Todd Dupont in their work "Interior Penalty Procedures for Elliptic and Parabolic Galerkin Methods" presented at the "Second International Symposium on Computing Methods in Applied Sciences and Engineering", held between December 15 and December 19 in 1975 organised by IRIA-LABORIA and then published in "Computing Methods in Applied Sciences", Springer-Verlag, Berlin, 1976, which contains part of the lectures which were presented during the mentioned conference. These node residuals are sometimes referred as Burman residuals because Erik Burman and Peter Hansbo, in their work "Edge stabilization for Galerkin approximations of convection-diffusion–reaction problems" published in 2004 in "Computer methods in applied mechanics and engineering", recalled the stabilization procedure introduced by Douglas and Dupon.

$$g|_{K^+}(\mathbf{x}) = \lim_{\substack{\mathbf{y}\to\mathbf{x} \\ \mathbf{y}\in K^+}} g(\mathbf{y})$$

for any $x \in f$;

- **Non-variational node residuals**

$$\mathbf{\Phi}_i^K(t, \mathbf{u}_h) = \beta_i^K(\mathbf{u}_h)\mathbf{\Phi}^K(t, \mathbf{u}_h) \quad t \in [0, T] \tag{6.8}$$

where $\beta_i^K(\mathbf{u}_h) \quad K \in \mathcal{T}_h \quad \mathbf{x}_i \in K$ are some coefficients depending on the approximated solution verifying

$$\sum_{\mathbf{x}_i \in K} \beta_i^K(\mathbf{u}_h) = 1 \quad \forall K \in \mathcal{T}_h. \tag{6.9}$$

Observation 37. *Due to the continuity of the basis functions φ_i we have a continuous approximation of the solution. For this reason no numerical flux is needed for the evaluation of the flux in the boundary integrals contrarily to what happens in the Discontinuous Galerkin case.*

Observation 38. *The Residual Distribution scheme obtained using Lax-Firedrichs node residuals is very robust anyway we will later show how with Lax-Firedrichs and SUPG node residuals we cannot reach more than first order accuracy. The Residual Distribution method with CIP node residuals is the reference one for the application of the Deferred Correction method and reach an explicit arbitrary high order formulation.*

Let's verify that the conservation relation 6.3 holds for all these types of node residuals. For what concerns the last type of node residuals this follows trivially from the condition 6.9 imposed on the coefficients $\beta_i^K(\mathbf{u}_h)$ in fact

$$\sum_{\mathbf{x}_i \in K} \mathbf{\Phi}_i^K(t, \mathbf{u}_h) = \sum_{\mathbf{x}_i \in K} \beta_i^K(\mathbf{u}_h)\mathbf{\Phi}^K(t, \mathbf{u}_h) = \left(\sum_{\mathbf{x}_i \in K} \beta_i^K(\mathbf{u}_h)\right) \mathbf{\Phi}^K(t, \mathbf{u}_h) = \mathbf{\Phi}^K(t, \mathbf{u}_h).$$

For the other node residuals let's refer to the next property.

Proposition 6. *The node residuals Lax-Friedrichs, SUPG and CIP given respectively in equations 6.5, 6.6 and 6.7 fulfil the conservation relation 6.3.*

Proof. For all of them it follows from the normalization 6.1 that we have made on the test functions which is recalled here

$$\sum_{i=1}^{I} \varphi_i(\mathbf{x}) \equiv 1 \quad \forall \mathbf{x} \in \overline{\Omega}.$$

In order to verify that the conservation relation holds let's observe that they all can be written in the form

$$\mathbf{\Phi}_i^K(t, \mathbf{u}_h) = \int_K \varphi_i(\mathbf{x}) \frac{\partial}{\partial t} \mathbf{u}_h(\mathbf{x}, t) d\mathbf{x} + \int_{\partial K} \varphi_i(\mathbf{x}) \mathbf{F}(\mathbf{u}_h(\mathbf{x}, t)) \cdot \nu d\sigma(\mathbf{x}) - \int_K \mathbf{F}(\mathbf{u}_h(\mathbf{x}, t)) \cdot \nabla_{\mathbf{x}} \varphi_i(\mathbf{x}) d\mathbf{x} +$$

$$-\int_K \mathbf{S}(\mathbf{x},t,\mathbf{u}_h(\mathbf{x},t))\varphi_i(\mathbf{x})d\mathbf{x} + \mathbf{ST}_i^K(\mathbf{u}_h) \quad t\in[0,T] \tag{6.10}$$

where $\mathbf{ST}_i^K(\mathbf{u}_h)$ is a stabilization term depending on the approximated solution $\mathbf{u}_h$. We have

- **Lax-Friedrichs**

$$\mathbf{ST}_i^K(\mathbf{u}_h) = \alpha_K(\mathbf{u}_h)\left(\mathbf{c}_i(t) - \overline{\mathbf{c}}_K(t)\right) \quad t\in[0,T];$$

- **SUPG**

$$\mathbf{ST}_i^K(\mathbf{u}_h) = h_K\int_K \left[\frac{\partial \mathbf{F}}{\partial \mathbf{u}}(\mathbf{u}_h(\mathbf{x},t))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})\right]\tau\left[\frac{\partial \mathbf{F}}{\partial \mathbf{u}}(\mathbf{u}_h(\mathbf{x},t))\cdot\nabla_{\mathbf{x}}\mathbf{u}_h(\mathbf{x},t)\right]d\mathbf{x} \quad t\in[0,T];$$

- **CIP**

$$\mathbf{ST}_i^K(\mathbf{u}_h) = \sum_{f\subset\partial K}\theta h_f^2\int_f [\![\nabla_{\mathbf{x}}\mathbf{u}_h(\mathbf{x},t)]\!]\cdot\nabla_{\mathbf{x}}\varphi_i|_K(\mathbf{x})d\sigma(\mathbf{x}) \quad t\in[0,T].$$

If we make now the sum of these nodal residuals associated to the nodes of a cell K we get

$$\sum_{\mathbf{x}_i\in K}\mathbf{\Phi}_i^K(t,\mathbf{u}_h) = \sum_{\mathbf{x}_i\in K}\Bigg[\int_K \varphi_i(\mathbf{x})\frac{\partial}{\partial t}\mathbf{u}_h(\mathbf{x},t)d\mathbf{x} + \int_{\partial K}\varphi_i(\mathbf{x})\mathbf{F}(\mathbf{u}_h(\mathbf{x},t))\cdot\nu d\sigma(\mathbf{x}) +$$

$$-\int_K \mathbf{F}(\mathbf{u}_h(\mathbf{x},t))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x} - \int_K \mathbf{S}(\mathbf{x},t,\mathbf{u}_h(\mathbf{x},t))\varphi_i(\mathbf{x})d\mathbf{x} + \mathbf{ST}_i^K(\mathbf{u}_h)\Bigg].$$

Due to the fact that we are dealing with a finite sum we can enter the sum under the integrals and, thanks to basic analysis, get

$$\sum_{\mathbf{x}_i\in K}\mathbf{\Phi}_i^K(t,\mathbf{u}_h) = \int_K\left(\sum_{\mathbf{x}_i\in K}\varphi_i(\mathbf{x})\right)\frac{\partial}{\partial t}\mathbf{u}_h(\mathbf{x},t)d\mathbf{x} + \int_{\partial K}\left(\sum_{\mathbf{x}_i\in K}\varphi_i(\mathbf{x})\right)\mathbf{F}(\mathbf{u}_h(\mathbf{x},t))\cdot\nu d\sigma(\mathbf{x}) +$$

$$-\int_K \mathbf{F}(\mathbf{u}_h(\mathbf{x},t))\cdot\nabla_{\mathbf{x}}\left(\sum_{\mathbf{x}_i\in K}\varphi_i(\mathbf{x})\right)d\mathbf{x} - \int_K \mathbf{S}(\mathbf{x},t,\mathbf{u}_h(\mathbf{x},t))\left(\sum_{\mathbf{x}_i\in K}\varphi_i(\mathbf{x})\right)d\mathbf{x} + \sum_{\mathbf{x}_i\in K}\mathbf{ST}_i^K(\mathbf{u}_h).$$

Now we have the crucial point: since, as we already said, each basis function φ_i has its support contained in the union of the elements containing the node to which it is associated then $\sum_{i=1}^I \varphi_i(\mathbf{x}) \equiv 1 \quad \forall \mathbf{x}\in\overline{\Omega}$ implies

$$\sum_{\mathbf{x}_i\in K}\varphi_i(\mathbf{x}) \equiv 1 \quad \forall x\in K$$

because only the basis functions φ_i s.t. the associated nodes $\mathbf{x}_i \in K$ are not identically zero in K. And obviously since the gradient of a constant function is identically 0 we can write

$$\nabla_{\mathbf{x}} \left(\sum_{\mathbf{x}_i \in K} \varphi_i(\mathbf{x}) \right) \equiv 0 \quad \forall x \in K. \tag{6.11}$$

Thus we get

$$\sum_{\mathbf{x}_i \in K} \mathbf{\Phi}_i^K(t, \mathbf{u}_h) = \int_K \frac{\partial}{\partial t} \mathbf{u}_h(\mathbf{x}, t) d\mathbf{x} + \int_{\partial K} \mathbf{F}(\mathbf{u}_h(\mathbf{x}, t)) \cdot \nu d\sigma(\mathbf{x}) - \int_K \mathbf{S}(\mathbf{x}, t, \mathbf{u}_h(\mathbf{x}, t)) d\mathbf{x} + \sum_{\mathbf{x}_i \in K} \mathbf{ST}_i^K(\mathbf{u}_h)$$

which, recalling the definition 6.2 of the element residuals, reads

$$\sum_{\mathbf{x}_i \in K} \mathbf{\Phi}_i^K(t, \mathbf{u}_h) = \mathbf{\Phi}^K(t, \mathbf{u}_h) + \sum_{\mathbf{x}_i \in K} \mathbf{ST}_i^K(\mathbf{u}_h) \quad t \in [0, T]. \tag{6.12}$$

All we need to prove now is that

$$\sum_{\mathbf{x}_i \in K} \mathbf{ST}_i^K(\mathbf{u}_h) \equiv \mathbf{0} \quad \forall t \in [0, T].$$

Thanks to basic analytical passages based on linearity and on the fact that we are dealing with finite sums we get

- **Lax-Friedrichs**

$$\sum_{\mathbf{x}_i \in K} \mathbf{ST}_i^K(\mathbf{u}_h) = \sum_{\mathbf{x}_i \in K} \alpha_K(\mathbf{u}_h) \left(\mathbf{c}_i(t) - \overline{\mathbf{c}}_K(t) \right) =$$

$$= \alpha_K(\mathbf{u}_h) \sum_{\mathbf{x}_i \in K} \left(\mathbf{c}_i(t) - \overline{\mathbf{c}}_K(t) \right) = \alpha_K(\mathbf{u}_h) \left(\sum_{\mathbf{x}_i \in K} \mathbf{c}_i(t) - \sum_{\mathbf{x}_i \in K} \overline{\mathbf{c}}_K(t) \right) =$$

$$= \alpha_K(\mathbf{u}_h) \left[\#\{\mathbf{x}_i \in K\} \overline{\mathbf{c}}_K(t) - \#\{\mathbf{x}_i \in K\} \overline{\mathbf{c}}_K(t) \right] \equiv 0 \quad \forall t \in [0, T]$$

because of the definition of the average $\overline{\mathbf{c}}_K(t) = \frac{1}{\#\{\mathbf{x}_i \in K\}} \sum_{\mathbf{x}_i \in K} \mathbf{c}_i(t)$;

- **SUPG**

$$\sum_{\mathbf{x}_i \in K} \mathbf{ST}_i^K(\mathbf{u}_h) = \sum_{\mathbf{x}_i \in K} h_K \int_K \left[\frac{\partial \mathbf{F}}{\partial \mathbf{u}}(\mathbf{u}_h(\mathbf{x}, t)) \cdot \nabla_{\mathbf{x}} \varphi_i(\mathbf{x}) \right] \tau \left[\frac{\partial \mathbf{F}}{\partial \mathbf{u}}(\mathbf{u}_h(\mathbf{x}, t)) \cdot \nabla_{\mathbf{x}} \mathbf{u}_h(\mathbf{x}, t) \right] d\mathbf{x} =$$

$$h_K \int_K \left[\frac{\partial \mathbf{F}}{\partial \mathbf{u}}(\mathbf{u}_h(\mathbf{x}, t)) \cdot \nabla_{\mathbf{x}} \left(\sum_{\mathbf{x}_i \in K} \varphi_i(\mathbf{x}) \right) \right] \tau \left[\frac{\partial \mathbf{F}}{\partial \mathbf{u}}(\mathbf{u}_h(\mathbf{x}, t)) \cdot \nabla_{\mathbf{x}} \mathbf{u}_h(\mathbf{x}, t) \right] d\mathbf{x} \equiv \mathbf{0} \quad \forall t \in [0, T]$$

because $\sum_{\mathbf{x}_i \in K} \varphi_i(\mathbf{x}) \equiv 1 \quad \forall x \in K$ and so its gradient is identically 0 in K as we observed in 6.11;

- **CIP**

$$\sum_{\mathbf{x}_i \in K} \mathbf{ST}_i^K(\mathbf{u}_h) = \sum_{\mathbf{x}_i \in K} \sum_{f \subset \partial K} \theta h_f^2 \int_f [\![\nabla_{\mathbf{x}} \mathbf{u}_h(\mathbf{x}, t)]\!] \cdot \nabla_{\mathbf{x}} \varphi_i|_K(\mathbf{x}) d\sigma(\mathbf{x}) =$$

$$= \sum_{f \subset \partial K} \theta h_f^2 \int_f [\![\nabla_{\mathbf{x}} \mathbf{u}_h(\mathbf{x}, t)]\!] \cdot \nabla_{\mathbf{x}} \left(\sum_{\mathbf{x}_i \in K} \varphi_i|_K(\mathbf{x}) \right) d\sigma(\mathbf{x}) \equiv \mathbf{0} \quad \forall t \in [0, T]$$

again because of 6.11 since we are considering the limit at the boundary of the restriction to K of the gradient of the function $\sum_{\mathbf{x}_i \in K} \varphi_i(\mathbf{x})$ which is constant in K and so

$$\nabla_{\mathbf{x}} \left(\sum_{\mathbf{x}_i \in K} \varphi_i|_K(\mathbf{x}) \right) = \nabla_{\mathbf{x}} \left(\sum_{\mathbf{x}_i \in K} \lim_{\substack{\mathbf{y} \to \mathbf{x} \\ \mathbf{y} \in K}} \varphi_i(\mathbf{y}) \right) =$$

$$= \nabla_{\mathbf{x}} \left(\lim_{\substack{\mathbf{y} \to \mathbf{x} \\ \mathbf{y} \in K}} \sum_{\mathbf{x}_i \in K} \varphi_i(\mathbf{y}) \right) = \nabla_{\mathbf{x}} 1 \equiv 0 \quad \forall x \in f \quad \forall f \subset \partial K.$$

□

6.1.2 A focus on the balance equations and on the mass matrix

We will now go deeper into the details of the Residual Distribution method. We will write explicitely the equations of the system of ODEs got from the balances at the nodes for the mentioned node residuals. Let's focus on the balance 6.4 that we recall here

$$\sum_{K \in K_i} \mathbf{\Phi}_i^K(t, \mathbf{u}_h) = \mathbf{0} \quad t \in [0, T] \quad \forall \imath = 1, 2, \ldots, I.$$

If we use Lax-Friedrichs, SUPG and CIP node residuals i.e. 6.5, 6.6 and 6.7, which as already shown can all be written in the form 6.10, the balance 6.4 at the nodes reads

$$\mathbf{0} = \sum_{K \in K_i} \mathbf{\Phi}_i^K(t, \mathbf{u}_h) = \sum_{K \in K_i} \Bigg[\int_K \varphi_i(\mathbf{x}) \frac{\partial}{\partial t} \mathbf{u}_h(\mathbf{x}, t) d\mathbf{x} + \int_{\partial K} \varphi_i(\mathbf{x}) \mathbf{F}(\mathbf{u}_h(\mathbf{x}, t)) \cdot \nu d\sigma(\mathbf{x}) +$$

$$- \int_K \mathbf{F}(\mathbf{u}_h(\mathbf{x}, t)) \cdot \nabla_{\mathbf{x}} \varphi_i(\mathbf{x}) d\mathbf{x} - \int_K \mathbf{S}(\mathbf{x}, t, \mathbf{u}_h(\mathbf{x}, t)) \varphi_i(\mathbf{x}) d\mathbf{x} + \mathbf{ST}_i^K(\mathbf{u}_h) \Bigg] \quad t \in [0, T] \quad \forall \imath = 1, 2, \ldots, I.$$

By substituiting $\mathbf{u}_h(\mathbf{x}, t) = \sum_{j=1}^I \mathbf{c}_j(t) \varphi_j(\mathbf{x})$ in the previous equation we get

$$\mathbf{0} = \sum_{K \in K_i} \Bigg[\int_K \varphi_i(\mathbf{x}) \frac{\partial}{\partial t} \left(\sum_{j=1}^I \mathbf{c}_j(t) \varphi_j(\mathbf{x}) \right) d\mathbf{x} + \int_{\partial K} \varphi_i(\mathbf{x}) \mathbf{F}(\sum_{j=1}^I \mathbf{c}_j(t) \varphi_j(\mathbf{x})) \cdot \nu d\sigma(\mathbf{x}) +$$

$$- \int_K \mathbf{F}(\sum_{j=1}^I \mathbf{c}_j(t) \varphi_j(\mathbf{x})) \cdot \nabla_{\mathbf{x}} \varphi_i(\mathbf{x}) d\mathbf{x} - \int_K \mathbf{S}(\mathbf{x}, t, \sum_{j=1}^I \mathbf{c}_j(t) \varphi_j(\mathbf{x})) \varphi_i(\mathbf{x}) d\mathbf{x} +$$

$$+\mathbf{ST}_i^K(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\Bigg] \quad t\in[0,T] \quad \forall \imath = 1,2,\ldots,I.$$

We remark that the basis function φ_j has support in the union of the elements containing the node $\mathbf{x}_j$ to which it is associated so the only basis functions having support in K are the ones corresponding to the nodes belonging to K thus we can write

$$\mathbf{0} = \sum_{K\in K_i}\Bigg[\int_K \varphi_i(\mathbf{x})\frac{\partial}{\partial t}\left(\sum_{\mathbf{x}_j\in K}\mathbf{c}_j(t)\varphi_j(\mathbf{x})\right)d\mathbf{x} + \int_{\partial K}\varphi_i(\mathbf{x})\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x}) +$$

$$-\int_K \mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x} - \int_K \mathbf{S}(\mathbf{x},t,\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\varphi_i(\mathbf{x})d\mathbf{x} +$$

$$+\mathbf{ST}_i^K(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\Bigg] \quad t\in[0,T] \quad \forall \imath = 1,2,\ldots,I.$$

Thanks to basic analytic passages the previous nodal balances become

$$\sum_{K\in K_i}\sum_{\mathbf{x}_j\in K}\left(\int_K \varphi_i(\mathbf{x})\varphi_j(\mathbf{x})d\mathbf{x}\right)\frac{d}{dt}\mathbf{c}_j(t) = -\sum_{K\in K_i}\Bigg[\int_{\partial K}\varphi_i(\mathbf{x})\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x}) +$$

$$-\int_K \mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x} - \int_K \mathbf{S}(\mathbf{x},t,\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\varphi_i(\mathbf{x})d\mathbf{x} +$$

$$+\mathbf{ST}_i^K(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\Bigg] \quad t\in[0,T] \quad \forall \imath = 1,2,\ldots,I. \tag{6.13}$$

This is the system of ODEs that we get with the Residual Distribution method and Lax-Friedrichs, SUPG and CIP node residuals. We remind that the stabilization terms $\mathbf{ST}_i^K$ depend on the particular node residuals chosen. The term at the left hand side of 6.13 "generates" the mass matrix $\mathcal{M}\in\mathbb{R}^{(I\times N)\times(I\times N)}$ which, in this case, is positive defined and thus nonsingular but unfortunately sparse[4].

Observation 39. *We remark that since we have a continuous approximation of the solution thanks to the continuous basis functions φ_i, we do not need a numerical flux for the evaluation of the flux in the boundary integrals of 6.13. The flux is a continuous function depending on the continuous appproximated solution.*

If we use insted the non-variational node residuals given by 6.8 we get

$$\mathbf{0} = \sum_{K\in K_i}\mathbf{\Phi}_i^K(t,\mathbf{u}_h) = \sum_{K\in K_i}\beta_i^K(\mathbf{u}_h)\mathbf{\Phi}^K(t,\mathbf{u}_h) =$$

[4]This means that we cannot rely on the efficient algorithms designed for band matrices as we can do instead in Discontinuous Galerkin FEM.

$$= \sum_{K\in K_i} \beta_i^K(\mathbf{u}_h)\left[\int_K \frac{\partial}{\partial t}\mathbf{u}_h(\mathbf{x},t)d\mathbf{x} + \int_{\partial K}\mathbf{F}(\mathbf{u}_h(\mathbf{x},t))\cdot\nu d\sigma(\mathbf{x}) + \right.$$

$$\left. - \int_K \mathbf{S}(\mathbf{x},t,\mathbf{u}_h(\mathbf{x},t))d\mathbf{x}\right] \quad t\in[0,T] \quad \forall \imath = 1,2,\ldots,I.$$

Substituiting again $\mathbf{u}_h(\mathbf{x},t) = \sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x})$ we have

$$\mathbf{0} = \sum_{K\in K_i} \beta_i^K(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\left[\int_K \frac{\partial}{\partial t}\left(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x})\right)d\mathbf{x} + \right.$$

$$\left. + \int_{\partial K}\mathbf{F}(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x}) - \int_K \mathbf{S}(\mathbf{x},t,\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))d\mathbf{x}\right] \quad t\in[0,T] \quad \forall \imath = 1,2,\ldots,I$$

and relying again on the fact that the only basis functions having support in K are the ones corresponding to the nodes belonging to K we get

$$\mathbf{0} = \sum_{K\in K_i} \beta_i^K(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\left[\int_K \frac{\partial}{\partial t}\left(\sum_{\mathbf{x}_j\in K} \mathbf{c}_j(t)\varphi_j(\mathbf{x})\right)d\mathbf{x} + \right.$$

$$\left. + \int_{\partial K}\mathbf{F}(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x}) - \int_K \mathbf{S}(\mathbf{x},t,\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))d\mathbf{x}\right] \quad t\in[0,T] \quad \forall \imath = 1,2,\ldots,I$$

Thanks to basic manipulations the previous node balances become

$$\sum_{K\in K_i}\sum_{\mathbf{x}_j\in K}\left[\beta_i^K(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\int_K \varphi_j(\mathbf{x})d\mathbf{x}\right]\frac{d}{dt}\mathbf{c}_j(t) =$$

$$= -\sum_{K\in K_i} \beta_i^K(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\left[\int_{\partial K}\mathbf{F}(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x}) + \right.$$

$$\left. - \int_K \mathbf{S}(\mathbf{x},t,\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))d\mathbf{x}\right] \quad t\in[0,T] \quad \forall \imath = 1,2,\ldots,I. \tag{6.14}$$

This is the system of ODEs that we get with the Residual Distribution method and non-variational node residuals. Also in this case the mass matrix is given by the term at the left hand side of the balance equation. In this case we do not even know if it is invertible or not and as, the reader may immediately notice, it depends on the coefficients $\mathbf{c}_i(t)$ through the coefficients β_i^K which take as argument the approximated solution $\mathbf{u}_h = \sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x})$. If in the first case avoiding to invert the mass matrix is just a matter of computational efficiency, now it is a must: even if we assume that it is nonsingular (which is not guaranteed), a traditional approach involving the inversion would imply at each time step to recompute (and invert) the mass matrix resulting in prohibitive computational times.

6.2 Stabilized continuous polynomial Galerkin FEM

We will will now introduce the Galerkin FEM for hyperbolic systems of balance laws. Basically it consists in a projection of the analytical weak formulation in space over a finite dimensional space. Let's thus derive the weak formulation in space by multiplying our initial equation 4.1 by a spatial test function with compact support and integrating in space

$$\int_\Omega \left(\frac{\partial}{\partial t}\mathbf{u}(\mathbf{x},t) + div_{\mathbf{x}}\mathbf{F}(\mathbf{u}(\mathbf{x},t)) \right) \varphi(\mathbf{x}) d\mathbf{x} = \int_\Omega \mathbf{S}(\mathbf{x},t,\mathbf{u}(\mathbf{x},t))\varphi(\mathbf{x}) d\mathbf{x} \quad \forall \varphi \in C_o^1(\overline{\Omega}).$$

If we apply the divergence theorem we get

$$\int_\Omega \varphi(\mathbf{x}) \frac{\partial}{\partial t}\mathbf{u}(\mathbf{x},t) d\mathbf{x} + \int_{\partial\Omega} \varphi(\mathbf{x}) \mathbf{F}(\mathbf{u}(\mathbf{x},t)) \cdot \nu d\sigma(\mathbf{x}) +$$

$$- \int_\Omega \mathbf{F}(\mathbf{u}(\mathbf{x},t)) \cdot \nabla_{\mathbf{x}} \varphi(\mathbf{x}) d\mathbf{x} = \int_\Omega \mathbf{S}(\mathbf{x},t,\mathbf{u}(\mathbf{x},t))\varphi(\mathbf{x},t) d\mathbf{x} \quad \forall \varphi \in C_o^1(\overline{\Omega}) \tag{6.15}$$

which is the weak formulation only in space. Now, according to the Galerkin FEM phylosophy, we project 6.15 over the finite dimensional subspace $(V_h)^N$. In particular, given a conformal tessellation $\mathcal{T}_h$ made of simplices[5], we adopt the usual discretization for the approximated solution

$$\mathbf{u}_h(\mathbf{x},t) = \sum_{i=1}^{I} \mathbf{c}_i(t)\varphi_i(\mathbf{x}) \quad \forall (\mathbf{x},t) \in \overline{\Omega} \times [0,T]$$

i.e. we assume at any time $t \in [0,T]$ the approximated solution $\mathbf{u}_h$ to belong to the finite dimensional space $(V_h)^N$ with $V_h = span\{\varphi_i\}_{i=1,2,\dots,I}$. The spatial basis functions φ_i are very general: they do not have to be mandatorily polynomials or continuous[6]. We will consider the same space V_h as the one considered in the Residual Distribution method i.e. the space of continuous, piecewise polynomial functions

$$V_h = \{ g \in C^0(\overline{\Omega}) \quad s.t. \quad g|_K \in \mathbb{P}_M \quad \forall K \in \mathcal{T}_h \}.$$

Let's remind that each function φ_i is associated to a node $\mathbf{x}_i \in \overline{\Omega}$ where its value is not 0 and has support in the union of the elements of the tessellation containing that node. Therefore as we anticipated we project 6.15 over the finite dimensional space $(V_h)^N$ i.e. we look for $\mathbf{u}_h(\mathbf{x},t) = \sum_{j=1}^{I} \mathbf{c}_j(t)\varphi_j(\mathbf{x}) \quad t \in [0,T]$ s.t.

$$\int_\Omega \varphi(\mathbf{x}) \frac{\partial}{\partial t}\mathbf{u}_h(\mathbf{x},t) d\mathbf{x} + \int_{\partial\Omega} \varphi(\mathbf{x}) \mathbf{F}(\mathbf{u}_h(\mathbf{x},t)) \cdot \nu d\sigma(\mathbf{x}) +$$

[5] Again we remark that this assumption is just meant to make things easier but it is not mandatory.

[6] If the basis functions are continuous we get a Continuous Galerkin FEM, otherwise we get a Discontinuous Galerkin FEM. Let's remind that when we deal with discontinuous approximations of the solution a numerical flux is needed when we have to evluate the flux in the boundary integrals.

$$-\int_\Omega \mathbf{F}(\mathbf{u}_h(\mathbf{x},t))\cdot\nabla_{\mathbf{x}}\varphi(\mathbf{x})d\mathbf{x}=\int_\Omega \mathbf{S}(\mathbf{x},t,\mathbf{u}_h(\mathbf{x},t))\varphi(\mathbf{x},t)d\mathbf{x}\quad\forall\varphi\in V_h.$$

Due to linearity this is equivalent to require

$$\int_\Omega \varphi_i(\mathbf{x})\frac{\partial}{\partial t}\mathbf{u}_h(\mathbf{x},t)d\mathbf{x}+\int_{\partial\Omega}\varphi_i(\mathbf{x})\mathbf{F}(\mathbf{u}_h(\mathbf{x},t))\cdot\nu d\sigma(\mathbf{x})+$$

$$-\int_\Omega \mathbf{F}(\mathbf{u}_h(\mathbf{x},t))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x}=\int_\Omega \mathbf{S}(\mathbf{x},t,\mathbf{u}_h(\mathbf{x},t))\varphi_i(\mathbf{x},t)d\mathbf{x}\quad\forall i=1,2,\ldots,I.$$

If we now substitute $\mathbf{u}_h(\mathbf{x},t)=\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x})$ in the previous equation we get

$$\int_\Omega \varphi_i(\mathbf{x})\frac{\partial}{\partial t}\left(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x})\right)d\mathbf{x}+\int_{\partial\Omega}\varphi_i(\mathbf{x})\mathbf{F}(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x})+$$

$$-\int_\Omega \mathbf{F}(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x}=\int_\Omega \mathbf{S}(\mathbf{x},t,\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\varphi_i(\mathbf{x})d\mathbf{x}\quad\forall j=1,2,\ldots,I.$$

and thus, thanks to basic analytic passages

$$\sum_{j=1}^I\left(\int_\Omega \varphi_i(\mathbf{x})\varphi_j(\mathbf{x})d\mathbf{x}\right)\frac{d}{dt}\mathbf{c}_j(t)=-\int_{\partial\Omega}\varphi_i(\mathbf{x})\mathbf{F}(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x})+$$

$$+\int_\Omega \mathbf{F}(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x}+\int_\Omega \mathbf{S}(\mathbf{x},t,\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\varphi_i(\mathbf{x})d\mathbf{x}\quad\forall i=1,2,\ldots,I. \tag{6.16}$$

This is the semidiscrete formulation of the Continuous Galerkin FEM method: a system of ODEs like 5.2 characterized by a mass matrix $\mathcal{M}\in\mathbb{R}^{(I\times N)\times(I\times N)}$ positive defined but sparse.

6.3 Link between Residual Distribution and stabilized continuous polynomial Galerkin FEM

We will now show that the system of ODEs given by 6.16 is equivalent to the one in 6.13 i.e. the system got with the Residual Distribution method and Lax-Friedrichs, SUPG and CIP node residuals up to the presence of the stabilization terms.

Since the basis function φ_i has support in the union of elements containing the node $\mathbf{x}_i$ to which it is associated, i.e. it is not identically zero just in the elements $K\in K_i$, we can consider in place of the integral over Ω the integral over $\cup_{K\in K_i}K$ and get

$$\sum_{j=1}^I\left(\int_{\cup_{K\in K_i}K}\varphi_i(\mathbf{x})\varphi_j(\mathbf{x})d\mathbf{x}\right)\frac{d}{dt}\mathbf{c}_j(t)=-\int_{\partial(\cup_{K\in K_i}K)}\varphi_i(\mathbf{x})\mathbf{F}(\sum_{j=1}^I \mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x})+$$

$$+\int_{\cup_{K\in K_i}K}\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x}+\int_{\cup_{K\in K_i}K}\mathbf{S}(\mathbf{x},t,\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\varphi_i(\mathbf{x})d\mathbf{x}\quad\forall i=1,2,\ldots,I.$$

which is equivalent to say

$$\sum_{j=1}^{I}\left(\sum_{K\in K_i}\int_K\varphi_i(\mathbf{x})\varphi_j(\mathbf{x})d\mathbf{x}\right)\frac{d}{dt}\mathbf{c}_j(t)=-\int_{\partial(\cup_{K\in K_i}K)}\varphi_i(\mathbf{x})\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x})+$$

$$+\sum_{K\in K_i}\int_K\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x}+\sum_{K\in K_i}\int_K\mathbf{S}(\mathbf{x},t,\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\varphi_i(\mathbf{x})d\mathbf{x}\quad\forall i=1,2,\ldots,I.$$

Since we are dealing with finite sums it is possible to interchange the sums over the elements and over all the nodes at the left hand side and enter the sum over all the nodes into the integral

$$\sum_{K\in K_i}\left(\int_K\varphi_i(\mathbf{x})\sum_{j=1}^{I}\varphi_j(\mathbf{x})d\mathbf{x}\right)\frac{d}{dt}\mathbf{c}_j(t)=-\int_{\partial(\cup_{K\in K_i}K)}\varphi_i(\mathbf{x})\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x})+$$

$$+\sum_{K\in K_i}\int_K\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x}+\sum_{K\in K_i}\int_K\mathbf{S}(\mathbf{x},t,\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\varphi_i(\mathbf{x})d\mathbf{x}\quad\forall i=1,2,\ldots,I.$$

Now, thanks to the usual argument that only the basis functions associated to the degrees of freedom in K are nonzero in K we can write

$$\sum_{K\in K_i}\left(\int_K\varphi_i(\mathbf{x})\sum_{\mathbf{x}_j\in K}\varphi_j(\mathbf{x})d\mathbf{x}\right)\frac{d}{dt}\mathbf{c}_j(t)=-\int_{\partial(\cup_{K\in K_i}K)}\varphi_i(\mathbf{x})\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x})+$$

$$+\sum_{K\in K_i}\int_K\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x}+\sum_{K\in K_i}\int_K\mathbf{S}(\mathbf{x},t,\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\varphi_i(\mathbf{x})d\mathbf{x}\quad\forall i=1,2,\ldots,I$$

and get

$$\sum_{K\in K_i}\sum_{\mathbf{x}_j\in K}\left(\int_K\varphi_i(\mathbf{x})\varphi_j(\mathbf{x})d\mathbf{x}\right)\frac{d}{dt}\mathbf{c}_j(t)=-\int_{\partial(\cup_{K\in K_i}K)}\varphi_i(\mathbf{x})\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x})+$$

$$+\sum_{K\in K_i}\Big[\int_K\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x}+\mathbf{S}(\mathbf{x},t,\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\varphi_i(\mathbf{x})d\mathbf{x}\Big]\quad\forall i=1,2,\ldots,I.$$

In order to identify this system of equations corresponding to the semidiscrete formulation of the Continuous Galerkin FEM with the one in 6.13 corresponding to the semidiscrete formulation of the Residual Distribution method with Lax-Friedrichs, SUPG and

CIP node residuals (up to the presence of the stabilization terms) we only need to show that

$$\int_{\partial(\cup_{K\in K_i}K)} \varphi_i(\mathbf{x})\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x}) = \sum_{K\in K_i}\int_{\partial K}\varphi_i(\mathbf{x})\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x}).$$

In order to do this let's observe that the quantities inside the integrals are the same. Thus we have to identify the differences in the sets over which they are integrated. If only one element K contains $\mathbf{x}_i$ the equality holds trivially because we would have on both sides the integral over the boundary ∂K of that single element. Let's focus on the more complicated case in which $\mathbf{x}_i$ belongs to more than one element. For a clearer view of what we are going to say let's refer to the figure 6.5 in which a bidimensional example is shown, the arguments are the same for any arbitrary D number of dimensions.

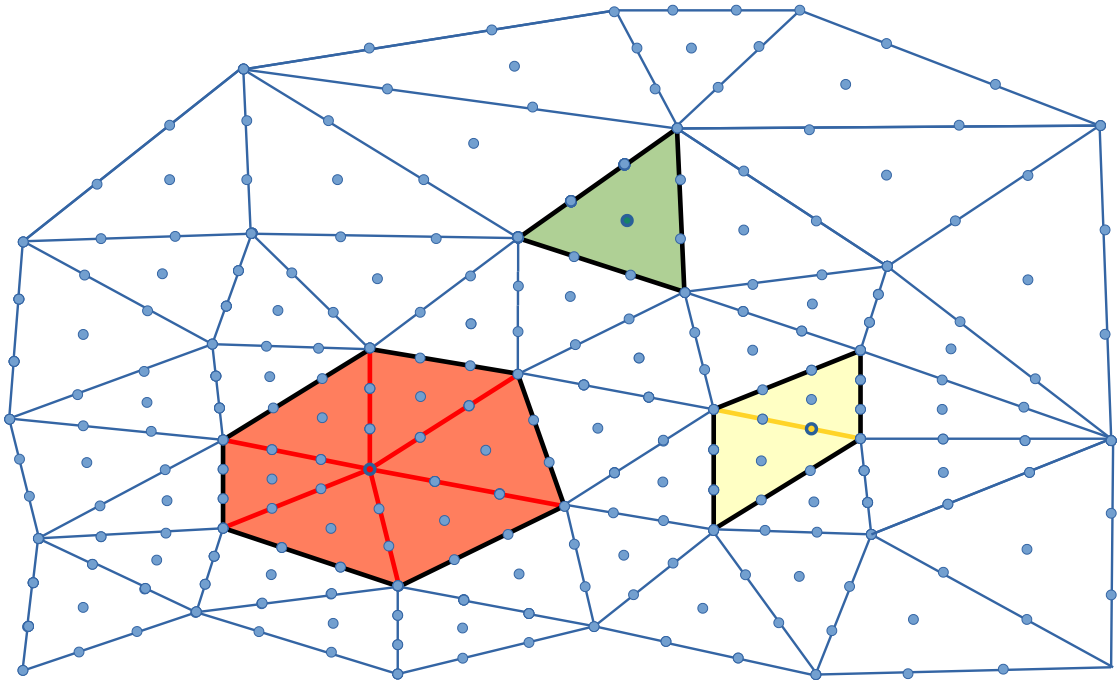


Figure 6.5: A focus on the boundaries of $\cup_{K\in K_i}K$ and $K \in K_i$.

Let's first focus on the left hand side of the equality. The set $\cup_{K\in K_i}K$ is a D-dimensional polytope resulting from the union of all the D-dimensional elements K containing $\mathbf{x}_i$. Under the assumption of conformal tessellation, its boundary $\partial(\cup_{K\in K_i}K)$ is made by the faces of the elements $K \in K_i$ which do not contain $\mathbf{x}_i$. This is basically due to the two facts:

- the node $\mathbf{x}_i$ is internal to $\cup_{K\in K_i}K$ and if a face of $K \in K_i$ contains $\mathbf{x}_i$ it means that it is internal to $\cup_{K\in K_i}K$,

- the elements of the tessellation are convex.

The faces of the elements $K \in K_i$ which do not contain $\mathbf{x}_i$ (whose union is $\partial(\cup_{K\in K_i} K)$) are in particular the faces of the elements $K \in K_i$ which are not shared with other elements in K_i. This comes as well from the convexity of the elements of the tessellation. In figure 6.5 three nodes are put in evidence with different colors: red, yellow and green. The elements K of the corresponding K_i are coloured with the same colour but with a lighter shade. The boundaries of the sets $\partial(\cup_{K\in K_i} K)$ are in black.

Let's now focus on the right hand side. We are integrating now over all the faces of K for each $K \in K_i$. The faces composing $\partial(\cup_{K\in K_i} K)$ are indeed counted among them but we are now considering also the faces containing $\mathbf{x}_i$. These are the faces shared between different elements $K \in K_i$. Due to the hypotesis of conformal tessellation each face is shared at most by two elements so in this case they are shared exactly by two simplices $K \in K_i$. So each of these shared faces is counted twice because we are considering the sum over all the elements K of K_i. In these two evaluations of the integral over the considered shared face we have opposite signs of the normal ν which is always outward pointing with respect to the considered element. This means that the two contributions cancel each other and this is equivalent to consider just the nonshared faces of the elements of K_i (which are obviously counted once). In figure 6.5 the shared (1-dimensional) faces of the elements $K \in K_i$ are of the same colour of the relative node $\mathbf{x}_i$. Obviously we have no shared faces in the naive case of the green node which belongs only to one element. We thus have that the equality is true and that the Residual Distribution method with Lax-Friedrichs, SUPG and CIP node residuals is equivalent to the Continuous Galerkin FEM up to the presence of the stabilization terms. In the context of these methods the stabilization terms are in general added to avoid the instabilities of the central schemes leading to the so-called "stabilized" methods[7]

We can conclude that the Residual Distribution method with Lax-Friedrichs, SUPG and CIP node residuals is equivalent to the stabilized Continuous Galerkin FEM. In particular we can put them in a common framework represented by the general system of ODEs

$$\sum_{K\in K_i}\sum_{\mathbf{x}_j\in K}\left(\int_K \varphi_i(\mathbf{x})\varphi_j(\mathbf{x})d\mathbf{x}\right)\frac{d}{dt}\mathbf{c}_j(t) = -\sum_{K\in K_i}\left[\int_{\partial K}\varphi_i(\mathbf{x})\mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nu d\sigma(\mathbf{x}) + \right.$$

$$\left. -\int_K \mathbf{F}(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\varphi_i(\mathbf{x})d\mathbf{x} - \int_K \mathbf{S}(\mathbf{x},t,\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x}))\varphi_i(\mathbf{x})d\mathbf{x}\right] +$$

$$+\mathbf{ET}_i(\sum_{j=1}^{I}\mathbf{c}_j(t)\varphi_j(\mathbf{x})) \quad t\in[0,T] \quad \forall_1 = 1,2,\ldots,I \tag{6.17}$$

[7]The stabilization term is usually added a priori in the weak formulation in space before the projection over a finite dimensional space through which we get the numerical method and not directly in the numerical method. This is done in order to derive some existence, regularity or stability results at the analytical level before passing to the numerical framework.

with $\mathbf{ET}_i$ being $\sum_{K \in K_i} \mathbf{ST}_i^K$ in the context of the Residual Distribution method with the three mentioned node residuals and a given stabilization term in the context of a stabilized Continuous Galerkin Finite Elements method. This will apply the Defect Correction method to this system of ODEs thus treating at the same time the Residual Distribution method with CIP residuals and the stabilized Continuous Galerkin Finite Elements method.

6.4 A little reference on the derivation of the stabilization term for the CIP nodal residuals

The name "CIP" stays for "continuous interior penalty" and is due to the fact that the Residual Distribution method that we get using these node residuals is equivalent to a stabilized Continuous Galerkin FEM with stabilization term based on the Interior Penalty Procedure introduced by Jim Douglas Jr. and Todd Dupont in their work "Interior Penalty Procedures for Elliptic and Parabolic Galerkin Methods" presented at the "Second International Symposium on Computing Methods in Applied Sciences and Engineering", held between December 15 and December 19 in 1975 organised by IRIA-LABORIA and then published in "Computing Methods in Applied Sciences", Springer-Verlag, Berlin, 1976, which contains part of the lectures which were presented during the mentioned conference. These node residuals are sometimes referred as Burman residuals because Erik Burman and Peter Hansbo, in their work "Edge stabilization for Galerkin approximations of convection-diffusion–reaction problems" published in 2004 in "Computer methods in applied mechanics and engineering", recalled the stabilization procedure introduced by Douglas and Dupon.

Chapter 7

The Deferred Correction in Residual Distribution and stabilized Continuous Galerkin FEM

Chapter 8

The Discontinuous Galerkin Finite Elements method

Chapter 9

The Deferred Correction in Discontinuous Galerkin

Chapter 10

The Deferred Correction as a Runge-Kutta scheme