Estimation of optimal penalization parameters and convergence of the non-linear solver for a Richards' equation-based model for variably-saturated groundwater flows based on an IIPG Discontinuous Galerkin method

Camille Poussel, Mehmet Ersoy, Frederic Golay, Damien Sous

▶ **To cite this version:**

Camille Poussel, Mehmet Ersoy, Frederic Golay, Damien Sous. Estimation of optimal penalization parameters and convergence of the non-linear solver for a Richards' equation-based model for variably-saturated groundwater flows based on an IIPG Discontinuous Galerkin method. 2024. hal-04704870

HAL Id: hal-04704870
https://hal.science/hal-04704870v1

Preprint submitted on 21 Sep 2024

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

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

Estimation of optimal penalization parameters and convergence of the non-linear solver for a Richards' equation-based model for variably-saturated groundwater flows based on an IIPG Discontinuous Galerkin method

C. Poussel, M. Ersoy, and F. Golay

Institut de Mathématiques de Toulon (IMATH), Université de Toulon, La Garde, France

D. Sous

Laboratoire des sciences pour l'ingénieur appliquées à la mécanique et au génie électrique (SIAME), Université de Pau et des Pays de l'Adour

September 21, 2024

Abstract

In this article, we study the convergence of an IIPG (Incomplete Interior Penalty Galerkin) Discontinuous Galerkin numerical method for the Richards equation. The Richards equation is a degenerate parabolic nonlinear equation for modeling flows in porous media with variable saturation. The numerical solution of this equation is known to be difficult to calculate numerically, due to the abrupt displacement of the wetting front, mainly as a result of highly non-linear hydraulic properties. As time scales are slow, implicit numerical methods are required and the convergence of nonlinear solvers is very sensitive. We propose an original method to ensure convergence of the numerical solution to the exact Richards solution, using a technique of auto-calibration of the penalty parameters derived from the Galerkin Discontinuous method. The method is constructed using non-linear 1D and 2D general elliptic problems. We show that the numerical solution converges toward the unique solution of the continuous problem under certain conditions on the penalty parameters. Then, we numerically demonstrate the efficiency and robustness of the method through test cases with analytical solutions, laboratory test cases, and large-scale simulations.

Keywords: Porous media, Richards Equation, Discontinuous Galerkin, Backward Differentiation Method, Incomplete Interior Penalty Galerkin (IIPG), Broken Soboloev space, Picard's fixed point, Minimal regularity solution

Contents

Acronyms

BDF Backward Differentiation Formula

DG Discontinuous Galerkin

FE Finite Element

IIPG Incomplete Interior Penalty Galerkin

ODE Ordinary Differential Equation

RE Richards' Equation

Introduction

The behavior of flows in variably-saturated porous media can be modeled by the Richards' Equation (RE). One of the key advantages of RE is its ability to represent the porous medium, incorporating both saturated and unsaturated zones. While it doesn't consider the air phase, RE effectively incorporates the effects of gravity and capillarity, enabling the modeling of complex processes across various scales. Notably, RE is a nonlinear parabolic equation that can transform into an elliptic equation under complete saturation conditions.

The history of RE begins with Darcy's law, which was formulated experimentally by Darcy in 1856 [12] for saturated porous media. This result was later extended to multiphase flows by Buckingham in 1907 [6], resulting in the Darcy-Buckingham law, which serves as the cornerstone for the derivation of RE. The equation was first established by Richardson in 1922 [33], although it was later attributed solely to Richards, who independently published the equation in 1931 [32]. Initial attempts to numerically solve the RE date back to the late 1960s with the works of Rubin [35] and Cooley [10]. From the 1980s, RE was extensively studied from both theoretical and numerical perspectives.

In this paper, RE is introduced by providing its expression and constitutive laws. As the main objective of this work is to solve RE using Discontinuous Galerkin (DG) methods, the weak problem associated with RE is given and its discretization using the Incomplete Interior Penalty Galerkin (IIPG) formulation. Additionally, an overview of the penalization method is provided. The fully discrete IIPG formulation is derived through time integration using the implicit Backward Differentiation Formula (BDF) method. Due to the non-linear nature of RE, its fully discretized non-linear formulation is linearized using the Picard's fixed point method. Theoretical results related to the solution of stationary non-linear elliptic problem are

produced, including existence, uniqueness, and convergence results. Furthermore, an automatic calibration method is obtained for penalization parameters. The solution of RE using the previously mentioned IIPG formulation is implemented in an in-house numerical code named `RIVAGE` which is then validated against numerical benchmarks.

1 Governing equation

RE is a classical nonlinear parabolic equation used to describe flow in both unsaturated and saturated zones of an aquifer (for a detailed derivation of the equation, please refer to Clement's 2021 thesis [8]).

The so-called mixed formulation of the RE, commonly used in hydrology, is

$$\partial_t \theta(h - z) - \nabla \cdot (\mathbb{K}(h - z)\nabla h) = 0 \tag{1}$$

where $h := \psi + z$ is the hydraulic head with ψ the pressure head, z the elevation, θ is the water content and $\mathbb{K}$ is the hydraulic conductivity tensor.

The tensor of hydraulic conductivity $\mathbb{K}$ is split, in general, into two parts, the intrinsic or saturated hydraulic conductivity tensor $\mathbb{K}_s$ and the relative hydraulic conductivity K_r:

$$\mathbb{K}(\psi) = \mathbb{K}_s K_r(\psi). \tag{2}$$

The intrinsic hydraulic conductivity tensor $\mathbb{K}_s$ depend on the material of the porous media.

The relative hydraulic conductivity is a function of the pressure head controlling the behavior of groundwater flow within the porous media and it is defined as

$$K_r(\psi) = \begin{cases} 1 & \text{if } \psi \geq \psi_e, \\ K_{e,\text{law}}(\psi) & \text{otherwise} \end{cases}$$

where $K_{e,\text{law}}$ is given by empirical laws, see Table 1 and Figure 1. The quantity ψ_e, corresponding to the entry of the air pressure, the pressure head transition value between the saturated and unsaturated zones. The saturated zone corresponds to $\psi \geq \psi_e$ and the unsaturated zone to $\psi < \psi_e$. The water table corresponds to $\psi = \psi_e$ by definition.

The water content law is expressed in terms of the effective saturation S_e:

$$S_e(\psi) = \frac{\theta(\psi) - \theta_r}{\theta_s - \theta_r}, \tag{3}$$

where θ_r is the residual water content and θ_s is the saturated water content corresponding to the minimal and maximal saturation, respectively. The effective saturation is defined as follows

$$S_e(\psi) = \begin{cases} 1 & \text{if } \psi \geq \psi_e, \\ S_{e,\text{law}}(\psi) & \text{otherwise,} \end{cases}$$

where $S_{e,\text{law}}$ is given by empirical laws, see Table 1 and Figure 1.

Remark 1.1. *The non-linear behavior of the constitutive laws $S_{e,law}$ and $K_{r,law}$ (see Table 1 and Figure 1) are responsible of the fails of the convergence of the numerical methods and a particular attention have been done. In particular, we have*

- *in the saturated zone, hydraulic properties remain constant and RE becomes an elliptic equation characterized by fast diffusion.*
- *in the unsaturated zone, hydraulic properties approach very close to zero, which halts diffusion and can cause numerical inconvenience.*

Name	**Expression**	**Parameters**
Gardner-Irmay relations (1958) [24]	$S_e = e^{\frac{\alpha\psi}{m}}$ $K_r = e^{\alpha\psi}$	α : pore-size distribution m : tortuosity
Vachaud's relations (1971) [41]	$S_e = \dfrac{C}{C + \lvert\psi\rvert^D}$ $K_r = \dfrac{A}{A + \lvert\psi\rvert^B}$	A, B : Empirical shape parameters C, D : Empirical shape parameters
Van Genuchen-Mualen relations (1980) [42]	$S_e = (1 + (\alpha\lvert\psi\rvert)^n)^{-m}$ $K_r = S_e^l \left(1 - \left(1 - S_e^{\frac{1}{m}}\right)^m\right)^2$	$l = 0.5$: pore connectivity α : linked to air entry pressure inverse $n > 1$: pore-size distribution $m = 1 - \dfrac{1}{n}$: pore-size distribution

Table 1: Hydraulic relations for hydraulic conductivity and effective saturation.

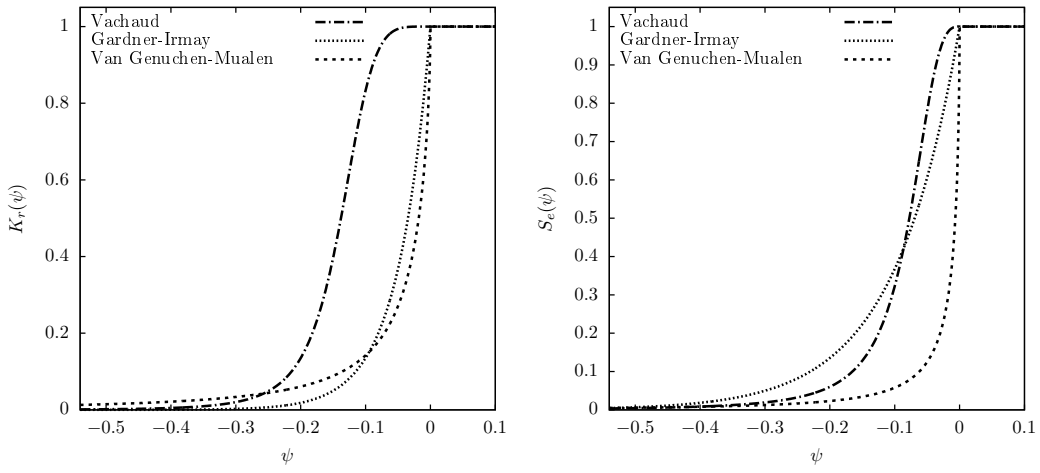


Figure 1: Hydraulic laws for effective saturation and hydraulic conductivity.

- *for a specific set of parameters, when $\psi \to 0^-$, constitutive laws may display extremely steep gradients.*

To overcome, regularization techniques can be employed as in [15], for instance, which make slight modifications to the functions to avoid some types of degeneracy to improve convergence properties. In this paper, we will see that in the framework of DG, we show that whenever some numerical parameters are well-chosen, the modification of such constitutive laws is not necessary.

Equation (1) together with Equation (2) and Equation (3) can be completed with Dirichlet and/or Neumann boundary conditions as done in this work. One can also use more realistic boundary condition in view of real life simulation, such as the seepage boundary condition (we refer to [9] for details).

2 Numerical methods

This section focus on the presentation of the numerical solution of RE using DG methods. The solution is sought within a trial space due to the similarity of these methods to Finite Element (FE) methods, resulting in a weak problem.

Let $d \in \{1, 2, 3\}$ be the space dimension, the porous medium can be represented by the computational domain $\Omega \subset \mathbb{R}^d$ of boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$ for which the subscript D and N stands for, respectively, Dirichlet and Neuman. Let $T \in \mathbb{R}_+^*$ be the final time.

The problem is:
Find $h(\mathbf{x}, t) : \Omega \times (0, T) \to \mathbb{R}$ such that:

$$\begin{cases} \partial_t \theta(h - z) - \nabla \cdot (\mathbb{K}(h - z)\nabla h) = 0 & \text{, in } \Omega \times (0, T), \\ h = h_0 & \text{, in } \Omega \times \{0\}, \\ h = h_D & \text{, on } \Gamma_D \times (0, T), \\ -\mathbb{K}(h - z)\nabla h \cdot \mathbf{n} = q_N & \text{, on } \Gamma_N \times (0, T) \end{cases} \tag{$\mathcal{P}_{\text{NL}}$}$$

where $h \in L^2(\Omega \times (0, T))$ represents the solution of RE. Additionally, $h_0 \in L^2(\Omega)$, $h_D \in L^2(\Gamma_D; (0, T))$, and $q_N \in L^2(\Gamma_N; (0, T))$ correspond to the initial condition, the Dirichlet boundary condition, and the Neumann boundary condition, respectively.

The matrix-valued function $\mathbb{K}$ depends monotonically on h, is symmetric positive definite, and is uniformly bounded below and above (see Equation (2), Table 1 and Figure 1). Similarly, the function θ, also depends monotonically on h, is uniformly bounded below and above (see Equation (3), Table 1 and Figure 1). Both $\mathbb{K}$ and θ are continuous functions within a given porous medium but may be discontinuous at the interface of heterogeneous materials.

2.1 Settings

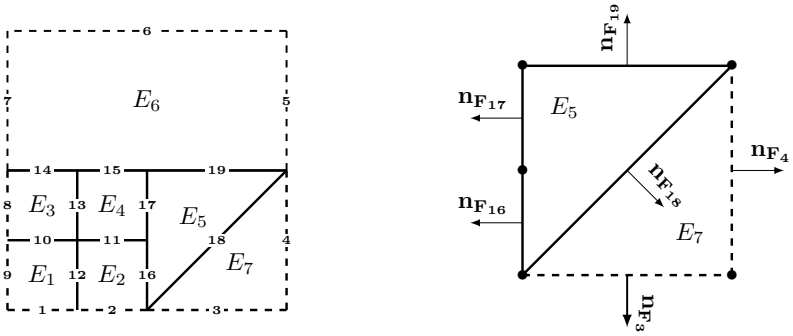
The time duration $(0, T)$ is subdivided into N time intervals such that $0 = t^0 < t^1 < \cdots < t^N = T$. Let $n \in \mathbb{N}$, $0 < n < N$, if the time interval $T^n = [t^n, t^{n+1}]$ is considered, the corresponding time step is $\Delta t^n = t^{n+1} - t^n$.

Let us define $\mathcal{E}^n$ a partition of the computational domain Ω valid for all $t \in T^n$. For the sake of simplicity, it is assumed that Ω is a polygonal domain in two space dimensions so that $\mathcal{E}^n$ covers Ω exactly. The mesh $\mathcal{E}^n$ is composed of quadrilateral and triangular elements not necessarily conformal.

For all elements $E \in \mathcal{E}^n$, d_E is its diameter defined as the ratio between its surface ($\mathfrak{s}_E$) and perimeter ($\mathfrak{p}_E$) and $d^n := \max_{E \in \mathcal{E}^n}(d_E)$.

The set of all open faces of all elements $E \in \mathcal{E}^n$ is denoted by $\mathcal{F}$. Moreover, one can define two subsets of $\mathcal{F}$, $\mathcal{F}^\partial$ for the boundary faces and $\mathcal{F}^{\text{in}}$ for the interior faces:

$$\mathcal{F}^\partial := \bigcup_{F \in \partial\Omega} F \quad \text{and} \quad \mathcal{F}^{\text{in}} := \mathcal{F} \backslash \mathcal{F}^\partial.$$



(a) Representation of $\mathcal{E}^n$, $\mathcal{F}^\partial$ (dashed lines) and $\mathcal{F}^{\text{in}}$ (solid lines)

(b) Description of E_5 and E_7 and their normal vectors

Figure 2: Exemple of a mesh.

For a given element $E \in \mathcal{E}^n$, there exists a set of face $\mathcal{F}^E := \{F \in \mathcal{F} | F \in \partial E\}$ which defines boundaries of E. Then for all interior faces of E, i.e. $\forall F \in \mathcal{F}^E \cap \mathcal{F}^{\text{in}}$, there exists a neighboring element E_r such that $E \cap E_r = F$. Consequently the normal unit vector $\mathbf{n_F} := (n_x, n_y)^T$ pointing from E to E_r can be defined. An exemple of interior face is given Figure 2a. Moreover for all boundary faces of E, i.e. $\forall F \in \mathcal{F}^E \cap \mathcal{F}^\partial$, there exists E_∂ a fictitious element such that $E \cap E_\partial = F$. Consequently, the normal unit vector $\mathbf{n_F}$ pointing always from E to E_∂ can be defined.

Example 2.1. *Figure 2a gives a graphical representation for an example mesh composed of triangles and quadrilaterals. In this exemple the mesh is composed of 7 elements, i.e.* $\mathcal{E}^n = \{E_i,\ i \in 1, \dots, 7\}$. *Thus the set of faces* $\mathcal{F} = \{F_i,\ i \in 1, \dots, 19\}$ *is defined. It can be split into two subsets, the first one* $\mathcal{F}^\partial = \{F_i,\ i \in 1, \dots, 9\}$ *boundary faces of* $\mathcal{F}$*, depicted with dashed lines on Figure 2. The second one* $\mathcal{F}^{in} = \{F_i,\ i \in 10, \dots, 19\}$ *interior faces of* $\mathcal{F}$*. Figure 2b gives graphical representation for two elements* E_5 *and* E_7*. Faces are also depicted with their normal vectors.*

Let two neighbouring elements E_l and E_r sharing one face $F \in \mathcal{F}$. There are two traces of a function v on E_l (v_l) and on E_r (v_r):

$$v_l(\mathbf{x}) := \lim_{\varepsilon \to 0^-} v(\mathbf{x} + \varepsilon \mathbf{n}_F) \text{ and } v_r(\mathbf{x}) := \lim_{\varepsilon \to 0^+} v(\mathbf{x} + \varepsilon \mathbf{n}_F),\ \forall \mathbf{x} \in F.$$

In addition, on any boundary faces $F \in \mathcal{F}^\partial$ the trace of v is only defined on the left side of the face:

$$v_l(\mathbf{x}) := \lim_{\varepsilon \to 0^-} v(\mathbf{x} + \varepsilon \mathbf{n}_F),\ \forall \mathbf{x} \in F$$

Using these trace definitions, one can define the jump and the average on any face of the mesh (as displayed in 1D on Figure 3). On an interior face $F \in \mathcal{F}^{\text{in}}$, the jump and the average are respectively defined as:

$$\forall \mathbf{x} \in F,\ [\![v]\!](\mathbf{x}) := v_r(\mathbf{x}) - v_l(\mathbf{x}) \text{ and } \{\!\{v\}\!\}(\mathbf{x}) := \frac{1}{2}\big(v_r(\mathbf{x}) + v_l(\mathbf{x})\big).$$

Moreover, on a boundary face $F \in \mathcal{F}^\partial$, the jump and the average are respectively defined as:

$$\forall \mathbf{x} \in F,\ [\![v]\!](\mathbf{x}) := v_l(\mathbf{x}) \text{ and } \{\!\{v\}\!\}(\mathbf{x}) := v_l(\mathbf{x}).$$

The solution of Problem ($\mathcal{P}_{\text{NL}}$) is sought in a subspace of the well-known broken Sobolev space, taken to be:

$$\mathcal{V}^p(\mathcal{E}^n) := \left\{ v \in L^2(\Omega) \;\middle|\; v|_E \in \mathbb{P}^p(E),\ \forall E \in \mathcal{E}^n \right\}$$

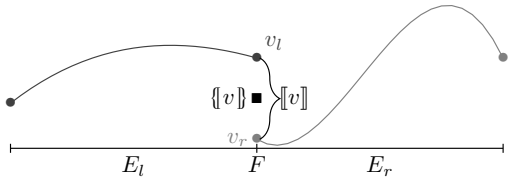


Figure 3: Definition of the mean and jump operators for two elements E_l and E_r in 1D.

where $\mathbb{P}^p(E)$ stands for the set of polynomial functions of degree less than or equal to $p \in \mathbb{N}$ on E. It is called the DG space. For more detailed and general definitions of this set, see [31].

2.2 Semi-discrete weak formulation

Keeping in mind that

$$\forall u, v \in \mathcal{V}^P(\mathcal{E}^n), \quad [\![uv]\!] = [\![u]\!]\{\!\{v\}\!\} + \{\!\{u\}\!\}[\![v]\!],$$

assuming that the flux of RE is continuous at the interfaces of elements:

$$\forall F \in \mathcal{F},\ [\![\mathbb{K}(h-z)\nabla h \cdot \mathbf{n_F}]\!]|_F = 0,$$

the Neumann boundary condition arises naturally in the weak formulation, multiplying Problem $(\mathcal{P}_{\mathrm{NL}})$ by a test function $\varphi \in \mathcal{V}^p(\mathcal{E}^n)$ and integrating on each element of $\mathcal{E}$, we get

$$\begin{cases} \displaystyle\sum_{E\in\mathcal{E}} \int_E \partial_t \theta(h-z)\varphi \mathrm{d}E + \sum_{E\in\mathcal{E}} \int_E (\mathbb{K}(h-z)\nabla h)\cdot\nabla\varphi \mathrm{d}E \\ \displaystyle -\sum_{F\in\mathcal{F}^{\mathrm{in}}} \int_F \{\!\{(\mathbb{K}(h-z)\nabla h)\cdot\mathbf{n_F}\}\!\}[\![\varphi]\!]\mathrm{d}F - \sum_{F\in\mathcal{F}^D}\int_F (\mathbb{K}(h-z)\nabla h)\cdot\mathbf{n_F}\varphi \mathrm{d}F \\ \displaystyle +\sum_{F\in\mathcal{F}^N}\int_F q_N\varphi \mathrm{d}F = 0, & \text{, on } t\in(0,T) \\ \displaystyle\sum_{E\in\mathcal{E}}\int_E h\varphi\mathrm{d}E = \sum_{E\in\mathcal{E}}\int_E h_0\varphi\mathrm{d}E, \\ h = h_D & \text{, on } \Gamma_D\times(0,T)\ . \end{cases} \tag{4}$$

To enforce the continuity of the solution and the Dirichlet boundary condition, two penalty terms are added:

$$J_I(h,\varphi) := \sum_{F\in\mathcal{F}^{\mathrm{in}}} \frac{1}{2}\left(\frac{\sigma_E^{in}}{d_E} + \frac{\sigma_{E_r}^{in}}{d_{E_r}}\right)\int_F [\![h]\!][\![\varphi]\!]\mathrm{d}F \tag{5}$$

$$J_D(h,\varphi) := \sum_{F\in\mathcal{F}^D} \frac{\sigma_E^{\partial}}{d_E}\int_F (h-h_D)\varphi\mathrm{d}F \tag{6}$$

where, J_I represents the penalization terms that constrain the continuity of the solution on the interior of the domain, and, J_D for the Dirichlet boundary conditions. σ_E^{in} and σ_E^{∂} are the penalization parameters for the interior and for the Dirichlet boundary condition where, we recall that, d_E is the diameter of an element E.

Remark 2.1. *This method is known as the IIPG method [9, 34]. The role of these parameters is essential to ensure the convergence of the method and will be studied in Section 3 for the first time, up to our knowledge, in the non-linear case. The linear case has been dealt in [16].*

Using Equation (5) and Equation (6) in Equation (4), the semi-discrete non-linear weak formulation of Problem $(\mathcal{P}_{\mathrm{NL}})$ is, $\forall\ t \in T^n$,

$$\begin{cases} \text{Find } h \in \mathcal{V}^p(\mathcal{E}^n) \text{ such that } : \\ m_n(\partial_t \theta(h-z), \varphi) + a_n(h, \varphi; h) = l_n(\varphi), \quad \forall \varphi \in \mathcal{V}^p(\mathcal{E}^n), \end{cases} \tag{$\mathcal{P}_{\mathrm{NLSD}}$}$$

where m_n, a_n, and l_n are given by:

$$m_n(q, \varphi) = \sum_{E \in \mathcal{E}^n} \int_E q\varphi \mathrm{d}E \tag{7}$$

$$\begin{aligned} a_n(h, \varphi; h) = & \sum_{E \in \mathcal{E}^n} \int_E (\mathbb{K}(h-z)\nabla h) \cdot \nabla \varphi \mathrm{d}E \\ & - \sum_{F \in \mathcal{F}^{\mathrm{in}}} \int_F \{\!\!\{ (\mathbb{K}(h-z)\nabla h) \cdot \mathbf{n_F} \}\!\!\} [\![\varphi]\!] \mathrm{d}F + \sum_{F \in \mathcal{F}^{\mathrm{in}}} \frac{1}{2} \left(\frac{\sigma_E^{in}}{d_E} + \frac{\sigma_{E_r}^{in}}{d_{E_r}} \right) \int_F [\![h]\!][\![\varphi]\!] \mathrm{d}F \\ & - \sum_{F \in \mathcal{F}^D} \int_F (\mathbb{K}(h-z)\nabla h) \cdot \mathbf{n_F} \varphi \mathrm{d}F + \sum_{F \in \mathcal{F}^D} \frac{\sigma_E^{\partial}}{d_E} \int_F h\varphi \mathrm{d}F \end{aligned} \tag{8}$$

$$l_n(\varphi) = \sum_{F \in \mathcal{F}^D} \frac{\sigma_E^{\partial}}{d_E} \int_F h_D \varphi \mathrm{d}F - \sum_{F \in \mathcal{F}^N} \int_F q_N \varphi \mathrm{d}F. \tag{9}$$

2.3 Time discretization

The aim of this section is to present the time discretisation through the implicit BDF method for Problem $(\mathcal{P}_{\mathrm{NLSD}})$. In the following, we make use of notation: $\forall n \in \mathbb{N}, u^n(\mathbf{x}) := u(\mathbf{x}, t_n)$, for any function $u \in L^2(\Omega \times (0,T))$. Let us recall that the time step is defined by $\Delta t^n = t^{n+1} - t^n$ and the time interval by $T^n = [t^n, t^{n+1}]$.

Due to their stability properties, the BDF methods are commonly used to solve stiff differential equations such as Problem $(\mathcal{P}_{\mathrm{NL}})$. These linear multi-step methods allows to construct time approximation up to order $q \leq 6$. The analysis of these methods can be found in [38]. The 1-step BDF method corresponds to the classical backward Euler scheme. BDF methods have been used in [26, 18] up to 6th-order. BDF methods are well-known to balance space and time errors and particularly well-designed in combination with DG methods. BDF methods can constructed both with a constant time step [38] or a variable [23]. The case of variable time step is more pertinent for Problem $(\mathcal{P}_{\mathrm{NLSD}})$ concerned. The method of order q is derived from the Newton interpolation polynomial of degree q, which interpolates h^j at time t^j for $j = n+1, \ldots, n+1-q$, using the method of divided difference.

The backward divided difference for a given function y is defined by a recursive division process:

$$\begin{cases} \delta^0 y^{n+1} = [y^{n+1}] = y^{n+1}, \\ \delta^1 y^{n+1} = [y^{n+1}, y^n] = \dfrac{\delta^0 y^{n+1} - \delta^0 y^n}{\Delta t^n} = \dfrac{y^{n+1} - y^n}{\Delta t^n}, \\ \delta^2 y^{n+1} = [y^{n+1}, y^n, y^{n-1}] = \dfrac{\delta^1 y^{n+1} - \delta^1 y^n}{\Delta t^n + \Delta t^{n-1}} = \dfrac{\frac{y^{n+1}-y^n}{\Delta t^n} - \frac{y^n - y^{n-1}}{\Delta t^{n-1}}}{\Delta t^n + \Delta t^{n-1}}, \\ \quad \vdots \\ \delta^j y^{n+1} = [y^{n+1}, y^n, \ldots, y^{n+1-j}] = \dfrac{\delta^{j-1} y^{n+1} - \delta^{j-1} y^n}{\sum_{k=0}^{j-1} \Delta t^{n-k}}. \end{cases}$$

For a given Ordinary Differential Equation (ODE), for instance $\frac{du}{dt} = f(u,t)$ with initial condition, the

Order q	**1**	**2**	**3**	**4**	**5**	**6**
Maximum swing $\Delta t^{n+2}/\Delta t^{n+1}$	–	2.6	1.9	1.5	1.2	1.05

Table 2: Maximum swing $\Delta t^{n+2}/\Delta t^{n+1}$ for BDF methods with variable time steps.

implicit BDF method of order q is given by:

$$\sum_{j=1}^{q}\left(\prod_{k=1}^{j-1}\left(\sum_{l=0}^{k-1}\Delta t^{n-l}\right)\right)\delta^j u^{n+1} = \sum_{j=0}^{q}\alpha_{q,j}u^{n+1-j} = f(u^{n+1},t^{n+1}),$$
$$\iff \alpha_{q,0}u^{n+1} - f(u^{n+1},t^{n+1}) = -\sum_{j=0}^{q-1}\alpha_{q,j+1}u^{n-j}$$

where $\alpha_{q,j}$ are the linear combination coefficients obtained from the divided differences of u. For instance, for the 2-order BDF method, the coefficients are:

$$\begin{cases}\alpha_{2,0} = \dfrac{1}{\Delta t^n} + \dfrac{1}{\Delta t^n + \Delta t^{n-1}},\\ \alpha_{2,1} = -\dfrac{1}{\Delta t^n} - \dfrac{1}{\Delta t^n + \Delta t^{n-1}} - \dfrac{\Delta t^n}{\Delta t^{n-1}(\Delta t^n + \Delta t^{n-1})},\\ \alpha_{2,2} = \dfrac{\Delta t^n}{\Delta t^{n-1}(\Delta t^n + \Delta t^{n-1})}.\end{cases}$$

Remark 2.2 (Stability)**.** *BDF methods of order* 1 *and* 2 *are A-stable, and L-stable [11]. BDF methods of order* 3 *to* 6 *are $A(\alpha)$-stable where α decreases as the order increases [21]. BDF methods of order $q > 6$ are unconditionally unstable. The use of variable time steps is recommended to enhance the stability of the method. In practical applications, variations in time step sizes are limited by an upper bound known as the swing factor to ensure stability and robustness Table 2 (see [36]). In the following, swing factors are used.*

Applying the BDF method to Problem ($\mathcal{P}_{\text{NLSD}}$), we get

$$\begin{cases}\text{Find a sequence of } (h^n)_{0\le n\le N} \in \mathcal{V}^p(\mathcal{E}^n) \text{ such that :}\\ m_n\left(\left.\dfrac{\partial\theta(\psi)}{\partial\psi}\right|_{\psi^{n+1}}\sum_{j=0}^{q}\alpha_{q,j}h^{n+1-j},\varphi\right) + a_n(h^{n+1},\varphi;h^{n+1}) = L_n(\varphi), \quad \forall\varphi\in\mathcal{V}^p(\mathcal{E}^n).\end{cases} \tag{$\mathcal{P}_{\text{NLFD}}$}$$

where m_n, a_n and l_n are given, respectively, by Equation (7), Equation (8), Equation (9) with $\psi = h - z$.

The time integration method needs an initialization step to compute the solution for further time steps. The initialization uses the prescribed initial condition to start the first time step. A direct and simple way is to write the corresponding discontinuous weak formulation:

$$\text{Find } h^0 \in \mathcal{V}^p(\mathcal{E}^0) \text{ such that: } m_0(h^0,\varphi) = f_0(\varphi),$$

where m_0 is defined by Equation (7) and f_0 is the linear form defined by:

$$f_0(\varphi) = \sum_{E\in\mathcal{E}^0}\int_E h_0\varphi \mathrm{d}E,\ \forall\varphi\in\mathcal{V}^p(\mathcal{E}^0).$$

2.4 Non-linear iterative process

Problem ($\mathcal{P}_{\text{NLFD}}$) being non-linear, several iterative methods can be used such as the Newton-Raphson method or the classical first-order fixed point method Picard's method. Due to the strong non-linearities of

the constitutive laws Equation (2) and Equation (3) (see also Remark 1.1), the convergence of the iterative methods may fails [25, 27]. We will see in Section 3 that in the case of IIPG methods one can enhance the convergence of the iterative methods, at least in the case of a Picard's fixed point method, whenever the penalization terms Equation (5) and Equation (6) are well-chosen. Therefore, in what follows, we present the Picard's fixed point method for Problem ($\mathcal{P}_{\mathrm{NLFD}}$).

Linearization of Problem ($\mathcal{P}_{\mathrm{NLFD}}$) is done by a Picards' iterative procedure. For $k = 0, \ldots,$ the problem is:

$$\begin{cases} \text{For a given } h^{n+1,k} \in \mathcal{V}^p(\mathcal{E}^n), \text{ find } h^{n+1,k+1} \in \mathcal{V}^p(\mathcal{E}^n) \text{ such that }, \ \forall \varphi \in \mathcal{V}^p(\mathcal{E}^n) : \\ m_n \left(\left. \frac{\partial \theta(\psi)}{\partial \psi} \right|_{\psi^{n+1,k}} \alpha_{q,0} h^{n+1,k+1}, \varphi \right) + a_n(h^{n+1,k+1}, \varphi; h^{n+1,k}) = \\ l_n(\varphi) - m_n \left(\left. \frac{\partial \theta(\psi)}{\partial \psi} \right|_{\psi^{n+1,k}} \sum_{j=0}^{q-1} \alpha_{q,j+1} h^{n-j}, \varphi \right). \end{cases} \tag{$\mathcal{P}_{\mathrm{LFD}}$}$$

where m_n, a_n and l_n are given, respectively, by Equation (7), Equation (8), Equation (9) with $\psi = h - z$. h^{n-j} stands for the solution at the rank k of the iterative process.

The global algorithm of the Picard's fixed-point iteration, for a positive n, is:

1. Start with an initial guess $h^{n+1,0}$;
2. Compute the solution of Problem ($\mathcal{P}_{\mathrm{LFD}}$) with $h^{n+1,0}$ to get $h^{n+1,1}$;
3. Start again with $h^{n+1,1}$;
4. ...
5. Compute the solution of Problem ($\mathcal{P}_{\mathrm{LFD}}$) with $h^{n+1,k}$ to get $h^{n+1,k+1}$;
6. Start again with $h^{n+1,k+1}$ until the stopping criteria are satisfied;
7. Set $h^{n+1} = h^{n+1,k+1}$.

The stopping criterion is one important choice in determining accuracy for a non-linear iterative process. For RE, the stopping criterion can be specified in terms of absolute error for pressure head or water content between two successive iterations [9]. For this study, we have used: $\frac{\|r_n(h,\varphi)\|_{L^2(\Omega)}}{\|a_n(h,\varphi)\|_{L^2(\Omega)}} < \varepsilon_1$ and $\frac{\|\delta_k\|_{L^2(\Omega)}}{\|h^k\|_{L^2(\Omega)}} < \varepsilon_2$, where $\delta_k = h^k - h^{k-1}$ and $r_n(h,\varphi) = m_n(\partial_t h, \varphi; h) + a_n(h, \varphi; h) - l_n(\varphi)$. ε_1 and ε_2 are user-defined tolerances. These two criteria are relative and independent of the characteristic quantities of the problem.

2.5 Adaptive time stepping

Time adaptation is motivated by the convergence of the nonlinear solver. On one hand, transient simulations have difficulties to converge if the time step is too large but, on the other hand, shorter time steps mean more time steps and so, a longer computational time. That is the reason why time adaptation is very attractive and common for Richards' equation. Different strategies can be used to adjust the time step [19, 3, 29], either heuristic and mainly based on convergence performance of the nonlinear solver or rational and based on error control. The latter ones are generally more efficient but heuristic methods remains a relevant approach due to their simplicity.

In this study, the time step is adjusted heuristically based on the number of iterations N_{it} from the nonlinear solver, as discussed in [39, 3]. The size of the time step directly influences the convergence of the solver. The simulations start with a time step Δt^0, and subsequent time steps are calculated according to the following rule: the time step remains unchanged if convergence is achieved between m_{it} and M_{it} nonlinear iterations; it is increased by an amplification factor $\lambda_{amp} > 1$ if convergence is achieved in fewer than m_{it}

nonlinear iterations; and it is decreased by a reduction factor $\lambda_{red} < 1$ if convergence requires more than M_{it} nonlinear iterations. If convergence fails due to solver issues (poor initial guess, bad condition number) or exceeds a prescribed maximum bound W_{it}, the time step is recalculated using a reduced step size ($\lambda_{red} < 1$). The calculation of the next time step Δt^{n+1} from the previous one Δt^n follows this time-stepping scheme:

$$\begin{cases} \Delta t^{n+1} = \begin{cases} \lambda_{amp}\Delta t^n \text{ if } N_{it} \leq m_{it}, \\ \Delta t^n \text{ if } m_{it} < N_{it} \leq M_{it}, \\ \lambda_{red}\Delta t^n \text{ if } M_{it} < N_{it} \leq W_{it}, \end{cases} \\ \Delta t^n = \lambda_{red}\Delta t^n \text{ if } N_{it} > W_{it} \text{ or if the solver has failed (time step is started again),} \end{cases}$$

with N_{it} the number of nonlinear iterations.

Remark 2.3. *By studying the full-time-dependent problem, as done in Section 3 in the case of the steady problem, the time step can be adjusted automatically and this work is in progress.*

Remark 2.4. *In the numerical code* `RIVAGE` *, Adaptive Mesh Refinement can be also employed. We refer to [17, 20, 1, 9, 8] for more details.*

3 Theoretical study and estimation of the optimal penalization parameters

In this section, we present the main result of this work, namely, the way to get a convergent iterative scheme by constructing a robust method to compute automatically the penalization parameters (see Equation (5) and Equation (6)). This is achieved by studying the theoretical properties and convergence of the solution of the discrete problem Problem ($\mathcal{P}_{\text{NLFD}}$) to the mathematical problem Problem ($\mathcal{P}_{\text{NL}}$). To this end and for the sake of simplicity, we will consider a toy model similar to the stationary RE for which we study, as depicted in Section 3,

1. the existence and uniqueness of the weak solution to the non-linear problem in Section 3.2.
2. the existence and uniqueness of the weak solution to the discrete linearized problem in Section 3.3.
3. the method to compute optimal penalization parameters to ensure the convergence of the non-linear solver at the discrete level in Section 3.4.
4. the convergence of the discrete linearized weak problem to the continuous linearized weak problem in Section 3.5

Proofs of this section are given in Appendix A and can be easily extended to several space dimensions. However, since the computations are rather technical to get the optimal penalization parameters in the two-dimensional case, for the sake of completeness, the 2D case for the existence and uniqueness of the weak solution to the discrete linearized problem is considered in Section 3.3. We will see that the construction of the optimal penalization parameters is essentially based on the constants appearing in the discrete continuity and the discrete coercivity of the operator.

3.1 Toy model

Let us consider the following toy problem ($\mathcal{P}$) on the interval $\Omega = [a, b] \subset \mathbb{R}$:

$$\begin{array}{l} \text{For a given } f \in L^2(\Omega), \text{ find } u(x) : \Omega \longrightarrow \mathbb{R} \text{ such that :} \\ \begin{cases} -\left(A(x, u, u')\right)' = f & \text{, in } \Omega \\ u = 0 & \text{, on } \partial\Omega \end{cases} \end{array} \qquad (\mathcal{P})$$

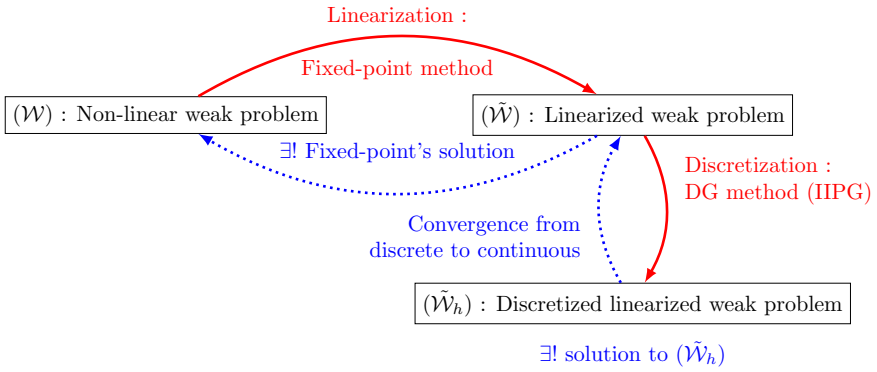


Figure 4: Scheme of the general proof.

with $A(x,s,\xi) = K(x,s)\xi$ where the real function K intends to mimick the properties of $\mathbb{K}$ (Equation (2)). Following [4] and in view of the properties of $\mathbb{K}$ (Equation (2)), assuming that

$$\begin{cases} \exists K_0, K_1 \in \mathbb{R}_+^* & , K_0 \leq K(x,\bar{u}) \leq K_1 & , \forall x \in \Omega,\ \forall \bar{u} \in \mathbb{R} \\ \exists K_{lip} \in \mathbb{R}_+ & , |K(x,\bar{u}_1) - K(x,\bar{u}_2)| \leq K_{lip}|\bar{u}_1 - \bar{u}_2| & , \forall x \in \Omega,\ \forall (\bar{u}_1,\bar{u}_2) \in \mathbb{R}^2 \end{cases}, \tag{$\mathcal{H}1$}$$

we deduce that A is straightforwardly a Carathéodory function, that we recall hereafter,

$$\begin{array}{lll} (1) & \exists \alpha > 0 & \text{s.t. } (A(x,s,\xi) - A(x,s,0))\xi \geq \alpha|\xi|^2, \\ (2) & \exists \beta > 0,\ \exists h \in L^2(\Omega) & \text{s.t. } |A(x,s,\xi)| \leq \beta(h(x) + |s| + |\xi|), \\ (3) & \exists \gamma > 0 & \text{s.t. } (A(x,s,\xi) - A(x,s,\eta))(\xi - \eta) \geq \gamma|\xi - \eta|^2, \\ (4) & \exists \delta > 0,\ \exists h \in L^2(\Omega) & \text{s.t. } |A(x,s,\xi) - A(x,t,\xi)| \leq \delta|s-t|(h(x) + |\xi| + |s| + |t|). \end{array} \tag{$\mathcal{H}2$}$$

This problem can be cast into the weak formulation by multiplying by a test function $v \in H_0^1(\Omega)$ and integrating over Ω:

$$\text{Find } u \in H_0^1(\Omega) \text{ such that : } a(u,v) = l(v),\ \forall v \in H_0^1(\Omega) \tag{$\mathcal{W}$}$$

where

$$a(u,v) = \int_\Omega K(x,u)u'v'\mathrm{d}x, \quad l(v) = \int_\Omega fv\mathrm{d}x.$$

Problem $(\mathcal{P})$ being non-linear, we use the Picard's iterations method as in Problem $(\mathcal{P}_{\text{NLFD}})$ to get

$$\begin{cases} \text{For a given } \bar{u} \in L^2(\Omega),\ \text{find } u \in H_0^1(\Omega) \text{ such that :} \\ \tilde{a}(u,v;\bar{u}) = l(v),\ \forall v \in H_0^1(\Omega) \end{cases} \tag{$\tilde{\mathcal{W}}$}$$

with

$$\tilde{a}(u,v;\bar{u}) = \int_\Omega K(x,\bar{u})u'v'dx\ .$$

Given $\bar{u}^0$, we solve the Problem $(\tilde{\mathcal{W}})$ with $\bar{u} = \bar{u}^0$ to obtain u^1. Then, we solve the Problem $(\tilde{\mathcal{W}})$ with $\bar{u} = \bar{u}^1$ to obtain u^2 and so on. The sequence of solutions of the linearized problem is denoted by $(u^n)_{n\in\mathbb{N}}$ and its limit when n goes to infinity is expected to be the solution to the non-linear Problem $(\mathcal{W})$. In the following we note $u^{n+1} = T(u^n)$ the fixed point.

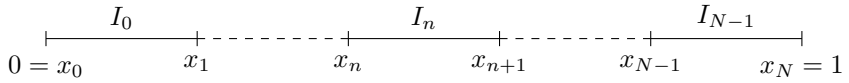


Figure 5: Representation of $\mathcal{E}_h$ in the one-dimension case.

3.2 Existence and uniqueness of the weak solution to the non-linear Problem $(\mathcal{W})$

The first step is to show that Problem $(\mathcal{W})$ has a unique solution in $H_0^1(\Omega)$. The existence of solution of Problem $(\mathcal{W})$ can be achieved by using the Schauder fixed-point theorem to the operator T while the uniqueness can be obtained through the technique proposed in [4].

Thus, we have

Lemma 3.1 (Existence of a solution to Problem $(\mathcal{W})$)**.** *Under Hypothesis $(\mathcal{H}1)$, $\exists u \in H_0^1(\Omega); T(u) = u$.*

Then, one can obtain uniqueness through the following result

Lemma 3.2 (Uniqueness of the solution to Problem $(\mathcal{W})$)**.** *Under Hypothesis $(\mathcal{H}1)$, the solution $u \in H_0^1(\Omega)$ of Problem $(\mathcal{W})$ is unique.*

These results hold for the dimension $d \leqslant 3$ and the proofs are rather classical and left to the reader.

3.3 Existence and uniqueness of the weak solution to the discrete linearized Problem $(\tilde{\mathcal{W}})$

One-dimensional case

To solve numerically Problem $(\tilde{\mathcal{W}})$, we use DG methods as in Section 2. Let $0 = x_0 < ... < x_N = 1$ be a partition $\mathcal{E}_h$ of Ω and denote $I_n = [x_n, x_{n+1}]$ a sub-interval. The size of a sub-interval is defined as $|I_n| := h = \frac{1}{N}, \quad \forall n \in \{0, .., N-1\}$ with N the number of elements in the partition. The solution is sought in the DG space $\mathcal{V}_0^p(\mathcal{E}_h)$ defined as:

$$\mathcal{V}_0^p(\mathcal{E}_h) = \left\{ v \in L^2(\Omega) \;\middle|\; v|_{\partial\Omega} = 0;\ v|_{I_n} \in \mathbb{P}^p(I_n),\ \forall I_n \in \mathcal{E}_h \right\} \subseteq L^2(\Omega)$$

As in Section 2, we define

$$v(x_n^+) = \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} v(x_n + \epsilon), \quad v(x_n^-) = \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} v(x_n - \epsilon),$$

$$[\![v]\!]_{x_n} = v(x_n^-) - v(x_n^+), \quad \{\!\{v\}\!\}_{x_n} = \frac{1}{2}\left(v(x_n^-) + v(x_n^+)\right), \quad \forall n \in \{1..N-1\},$$

and

$$[\![v]\!]_{x_0} = -v(x_0^+),\ \{\!\{v\}\!\}_{x_0} = v(x_0^+), \quad [\![v]\!]_{x_N} = v(x_N^-),\ \{\!\{v\}\!\}_{x_N} = v(x_N^-).$$

The DG space $\mathcal{V}_0^p(\mathcal{E}_h)$ is associated with the norm:

$$\|v\|^2 = \sum_{n=0}^{N-1} \|v'\|_{I_n}^2 + \sum_{n=0}^{N} \frac{1}{h} [\![v]\!]_{x_n}^2 = \sum_{n=0}^{N-1} \|v'\|_{I_n}^2 + |v|_J^2 \tag{10}$$

where $\|\cdot\|_{I_n}$ is the usual norm $L^2(I_n)$ and $|v|_J^2 := \sum_{n=0}^{N} \frac{1}{h} [\![v]\!]_{x_n}^2$ is the jump semi-norm. With this definition of the norm, jumps are controlled. One can observe that $\|\cdot\|$ is a norm on $\mathcal{V}_0^p(\mathcal{E}_h)$. One can note that $\mathcal{V}_0^p(\mathcal{E}_h)$ is a complete Banach space, i.e., a complete normed vector space for $\|\cdot\|$. Lastly the concept of

broken gradient is introduced to specify when only the regular part of the gradient is considered. The broken gradient $\nabla_h : \mathcal{V}_0^p(\mathcal{E}_h) \to L^2(\Omega)$ is defined such that, for all $v \in \mathcal{V}_0^p(\mathcal{E}_h)$,

$$\forall E \in \mathcal{E}_h,\ (\nabla_h v)|_E := \nabla(v|_E).$$

The linearized weak formulation Problem $(\tilde{\mathcal{W}})$ can be discretized using the IIPG formulation as in Section 2 to get

$$\begin{cases} \text{For a given } \bar{u} \in \mathcal{V}_0^p(\mathcal{E}_h),\ \text{find } u_h \in \mathcal{V}_0^p(\mathcal{E}_h) \text{ such that :} \\ \tilde{a}_h(u_h, v_h; \bar{u}) = l_h(v_h),\ \forall v_h \in \mathcal{V}_0^p(\mathcal{E}_h) \end{cases} \tag{$\tilde{\mathcal{W}}_h$}$$

with

$$\begin{aligned} \tilde{a}_h(u_h, v_h, \bar{u}) = &\sum_{n=0}^{N-1} \int_{I_n} K(x, \bar{u}) u_h' v_h' dx - \sum_{n=0}^{N} \{\!\{ K(x, \bar{u}) u_h' \}\!\}_{x_n} [\![v_h]\!]_{x_n} \\ &+ \frac{\sigma_0}{h} [\![u_h]\!]_{x_0} [\![v_h]\!]_{x_0} + \sum_{n=1}^{N-1} \frac{\sigma_{n-1} + \sigma_n}{2h} [\![u_h]\!]_{x_n} [\![v_h]\!]_{x_n} + \frac{\sigma_N}{h} [\![u_h]\!]_{x_N} [\![v_h]\!]_{x_N}, \\ l_h(v_h) = &\int_\Omega f v_h dx. \end{aligned} \tag{11}$$

At the discrete level, one can write Hypothesis $(\mathcal{H}1)$ as follows: for all $n \in \{0, ..., N-1\}$:

$$\begin{cases} \exists K_0^{(n)}, K_1^{(n)} \in \mathbb{R}_+^*,\ \forall x \in I_n,\ \forall \bar{u} \in \mathbb{R}, & K_0^{(n)} \le K(x, \bar{u}) \le K_1^{(n)}; \\ \exists K_{lip}^{(n)} \in \mathbb{R}_+,\ \forall x \in I_n,\ \forall (\bar{u}_1, \bar{u}_2) \in \mathbb{R}^2, & |K(x, \bar{u}_1) - K(x, \bar{u}_2)| \le K_{lip} |\bar{u}_1 - \bar{u}_2| \end{cases} \tag{$\mathcal{H}_n$}$$

where

$$K_1 := \min_{n=0,...,N-1} K_1^{(n)},\ K_0 := \max_{n=0,...,N-1} K_0^{(n)} \text{ and } K_{lip} = \max_{n=0,...,N-1} K_{lip}^{(n)}.$$

Existence and unicity for the solution to Problem $(\tilde{\mathcal{W}}_h)$ is obtained using the Lax-Milgram theorem. We have the following result:

Theorem 3.1 (Existence and uniqueness of the weak solution to the discrete linearized Problem $(\tilde{\mathcal{W}}_h)$)**.** *Under Hypothesis $(\mathcal{H}_n)$ for all n, for a given $\bar{u} \in \mathcal{V}_0^p(\mathcal{E}_h)$, then $\exists! u \in \mathcal{V}_0^p(\mathcal{E}_h)$ such that $\tilde{a}_h(u_h, v_h; \bar{u}) = l_h(v_h),\ \forall v_h \in \mathcal{V}_0^p(\mathcal{E}_h)$.*

This existence and uniqueness result is obtained thanks to the below-following lemmas.

Lemma 3.3 (Discrete coercivity of $\tilde{a}_h$)**.** *Under Hypothesis $(\mathcal{H}_n)$ for all n, for any vector of positive numbers $\epsilon = (\varepsilon^{(n)})_{n=0,...,N-1}$, there exists a constant $C^*(\epsilon) > 0$ such that*

$$\forall u_h \in \mathcal{V}_0^p(\mathcal{E}_h),\ \tilde{a}_h(u_h, u_h; \bar{u}) \ge C^*(\epsilon) \|u_h\|^2$$

if

$$\begin{cases} \varepsilon^{(n)} < 2, & \forall n \in \{0, ..., N-1\} \\ \sigma_n > \sigma_n^*, & \forall n \in \{1, ..., N-1\} \\ \sigma_0 > \sigma_0^* \\ \sigma_N > \sigma_N^* \end{cases} \quad \textit{with} \quad \begin{cases} \sigma_n^* = \dfrac{(K_1^{(n)} C_{tr,p-1}^{(n)})^2}{2\varepsilon^{(n)} K_0^{(n)}},\ \forall n \in \{1, ..., N-1\} \\ \sigma_0^* = \dfrac{(K_1^{(0)} C_{tr,p-1}^{(0)})^2}{\varepsilon^{(0)} K_0^{(0)}} \\ \sigma_N^* = \dfrac{(K_1^{(N-1)} C_{tr,p-1}^{(N-1)})^2}{\varepsilon^{(N-1)} K_0^{(N-1)}} \end{cases}$$

and

$$C^*(\epsilon) = \min \Big\{ \min_{n=0,..,N-1} \Big(K_0^{(n)} (1 - \frac{\varepsilon^{(n)}}{2}) \Big), \sigma_0 - \sigma_0^*, \sigma_N - \sigma_N^*, \min_{n=1,..,N-1} \Big(\frac{\sigma_n - \sigma_n^*}{2} \Big) \Big\}.$$

Lemma 3.4 (Discrete continuity of $\tilde{a}_h$)**.** *Under Hypothesis ($\mathcal{H}_n$) for all n, for any vector of positive numbers $\epsilon = (\varepsilon^{(n)})_{n=0,\dots,N-1}$, there exists a constant $\tilde{C}(\epsilon) > 0$ such that*

$$\forall u_h, v_h \in \mathcal{V}_0^p(\mathcal{E}_h),\ |\tilde{a}_h(u_h, v_h; \bar{u})| \leq \tilde{C}(\epsilon)\|u_h\|\,\|v_h\|$$

where

$$\tilde{C}(\epsilon) = \max_{n=0,\dots,N-1}\left(K_1^{(n)}\right) + \sqrt{\max_{n=0,\dots,N-1}\left(2\varepsilon^{(n)}K_1^{(n)}\right)\max\left(\sigma_0^*, \sigma_N^*, \max_{n=1,\dots,N-1}\left(\frac{\sigma_n^*}{2}\right)\right)} + \max\left(\sigma_0, \sigma_N, \max_{n=1,\dots,N-1}\left(\frac{\sigma_n}{2}\right)\right).$$

Lemma 3.5 (Discrete continuity of l_h)**.** *There exists a constant $B > 0$ such that $\forall v_h \in \mathcal{V}_0^p(\mathcal{E}_h),\ \|l_h(v_h)\| \leq B\|v_h\|$.*

Remark 3.1. *Trace constant involved in bounds for penalization parameters are a function of the polynomial degree p, the type of polynomial basis used. In the one-dimensional case, with an orthonormal basis and for $u \in \mathcal{V}_0^p(\mathcal{E}_h)$, the trace constant for I_n is given by:*

$$C_{tr,p}^{(n)} := p + 1.$$

Proofs of Lemmas 3.3, 3.4, 3.5 can be found in Appendix A . The proof of Theorem 3.1 is a straightforward application of the Lax-Milgram theorem and is left to the reader.

Two-dimensional case

We propose to extend the previous results to the dimension 2. Let us consider the two-dimensional extension of Problem ($\tilde{\mathcal{W}}_h$)

$$\begin{cases}\text{For a given } \bar{u} \in \mathcal{V}_0^p(\mathcal{E}_h), \text{find } u_h \in \mathcal{V}_0^p(\mathcal{E}_h) \text{ such that },\ \forall v_h \in \mathcal{V}_0^p(\mathcal{E}_h) : \\ \tilde{a}_h(u_h, v_h; \bar{u}) = l_h(v_h).\end{cases} \tag{$\tilde{\mathcal{W}}_h^2$}$$

where

$$\begin{aligned}\tilde{a}(u_h, v_h; \bar{u}) = &\sum_{E\in\mathcal{E}^n}\int_E (\mathbb{K}(\mathbf{x}, \bar{u})\nabla u_h)\cdot\nabla v_h \mathrm{d}E - \sum_{F\in\mathcal{F}}\int_F \{\!\{(\mathbb{K}(\mathbf{x}, \bar{u})\nabla u_h)\cdot\mathbf{n_F}\}\!\}[\![v_h]\!]\mathrm{d}F \\ &+ \sum_{F\in\mathcal{F}^{\mathrm{in}}}\frac{1}{2}\left(\frac{\sigma_E^{in}}{d_E} + \frac{\sigma_{E_r}^{in}}{d_{E_r}}\right)\int_F [\![u_h]\!][\![v_h]\!]\mathrm{d}F + \sum_{F\in\mathcal{F}^D}\frac{\sigma_E^{\partial}}{d_E}\int_F u_h v_h \mathrm{d}F \\ l_h(v_h) = &\int_\Omega f v_h dx.\end{aligned}$$

The two-dimensional version of the discrete hypothesis on $\mathbb{K}$ is given by: For all $E \in \mathcal{E}$:

$$\left\{\exists K_0^E, K_1^E \in \mathbb{R}_+^*,\ \forall \mathbf{x} \in E,\ \forall \bar{u} \in \mathbb{R},\quad K_0^E \leq \|\mathbb{K}(\mathbf{x}, \bar{u})\|_2 \leq K_1^E; \right. \tag{$\mathcal{H}_E^2$}$$

with $\|\mathbb{K}\|_2 = \max_{i=1,2}(\mathbb{K}_{ii})$. In addition, $K_1 = \max_{E\in\mathcal{E}} K_1^E$ and $K_0 = \min_{E\in\mathcal{E}} K_0^E$ denotes global bound of $\mathbb{K}$.

The DG space is associated with the following norm:

$$\left\|v\right\|^2 := \sum_{E\in\mathcal{E}}\left\|v\right\|_E^2 + \sum_{F\in\mathcal{F}^{\mathrm{in}}}\left(\frac{1}{d_E} + \frac{1}{d_{E_r}}\right)\left\|[\![v]\!]\right\|_F^2 + \sum_{F\in\mathcal{F}^{\partial}}\frac{1}{d_E}\left\|[\![v]\!]\right\|_F^2 = \sum_{E\in\mathcal{E}}\left\|v\right\|_E^2 + |v|_J^2$$

where $\left\|v\right\|_E^2$ is the usual L^2 norm on E, $\left\|v\right\|_F^2$ is the L^2 norm on F and $|v|_J^2$ is the jump semi-norm. This norm has the same characteristics as in the one-dimensional case. We obtain the following result

Theorem 3.2 (Existence and uniqueness of the weak solution to the discrete linearized Problem $(\tilde{\mathcal{W}}_h^2)$)**.** *If K satisfies Hypothesis $(\mathcal{H}_E^2)$ for all $E \in \mathcal{E}$ and for a given $\bar{u} \in \mathcal{V}_0^p(\mathcal{E}_h)$, then $\exists! u \in \mathcal{V}_0^p(\mathcal{E}_h)$ such that $\tilde{a}_h(u_h, v_h; \bar{u}) = l_h(v_h),\ \forall v_h \in \mathcal{V}_0^p(\mathcal{E}_h)$.*

As before, This result is a consequence of the Lax-Milgram theorem through the following lemmas:

Lemma 3.6 (Discrete coercivity of $\tilde{a}_h$)**.** *If K satisfies Hypothesis $(\mathcal{H}_E^2)$ for all $E \in \mathcal{E}$ and for any vector of positive numbers $\epsilon = (\varepsilon^E)_{E\in\mathcal{E}}$, there exists a constant $C^*(\epsilon) > 0$ such that*

$$\forall u_h \in \mathcal{V}_0^p(\mathcal{E}_h),\ \tilde{a}_h(u_h, u_h; \bar{u}) \geq C^*(\epsilon)\|u_h\|^2$$

if

$$\begin{cases} \varepsilon^E < 2, & \forall E \in \mathcal{E} \\ \sigma_E^{in} > \sigma_E^{in,*} \text{ and } \sigma_{E_r}^{in} > \sigma_{E_r}^{in,*}, & \forall F \in \mathcal{F}^{in} \\ \sigma_E^{\partial} > \sigma_E^{\partial,*}, & \forall F \in \mathcal{F}^{\partial} \end{cases} \text{ with } \forall E \in \mathcal{E} \begin{cases} \sigma_E^{in,*} = \dfrac{D^E \left(K_1^E C_{tr,p-1}^E\right)^2}{4\varepsilon^E K_0^E} \\ \sigma_E^{\partial,*} = \dfrac{D^E \left(K_1^E C_{tr,p-1}^E\right)^2}{2\varepsilon^E K_0^E} \end{cases}$$

and D^E is the number of edges of the element E. Moreover

$$C^*(\epsilon) = \min\left\{ \min_{E\in\mathcal{E}} \left(K_0^E(1 - \frac{\varepsilon^E}{2}) \right), \min_{E\in\mathcal{E}} \left(\frac{\sigma_E^{in} - \sigma_E^{in,*}}{2} \right), \min_{F\in\mathcal{F}^{\partial}} \left(\sigma_E^{\partial} - \sigma_E^{\partial,*} \right) \right\}.$$

Lemma 3.7 (Discrete continuity of $\tilde{a}_h$)**.** *If K satisfies Hypothesis $(\mathcal{H}_E^2)$ for all $E \in \mathcal{E}$ and for any vector of positive numbers $\epsilon = (\varepsilon^E)_{E\in\mathcal{E}}$, there exists a constant $\tilde{C}(\epsilon) > 0$ such that*

$$\forall u_h, v_h \in \mathcal{V}_0^p(\mathcal{E}_h),\ |\tilde{a}_h(u_h, v_h; \bar{u})| \leq \tilde{C}(\epsilon)\|u_h\|\,\|v_h\|$$

where

$$\begin{aligned} \tilde{C}(\epsilon) = & \max_{E\in\mathcal{E}} K_1^E + \max\left\{ \max_{E\in\mathcal{E}} \left(\frac{\sigma_E^{in}}{2} \right), \max_{F\in\mathcal{F}^{\partial}} \left(\sigma_E^{\partial} \right) \right\} \\ & + \sqrt{2 \max_{E\in\mathcal{E}} \varepsilon^E K_1^E \max\left\{ \max_{E\in\mathcal{E}} \left(\frac{\sigma_E^{in,*}}{2} \right), \max_{F\in\mathcal{F}^{\partial}} \left(\sigma_E^{\partial,*} \right) \right\}}. \end{aligned}$$

Lemma 3.5 still holds in the two-dimensional case and is left to the reader. Proofs of Lemmas 3.2, 3.6 and 3.7 are similar to proofs in the one-dimensional case. The main difference is in the expression of trace constants. In two dimensions, they are linked to the element's shape. For an orthonormal basis and for $u \in \mathcal{V}_0^p(\mathcal{E}_h)$, the trace constant of $E \in \mathcal{E}$ is given by:

$$C_{\mathrm{tr},p}^E = \begin{cases} \sqrt{\dfrac{(p+1)(p+2)}{2}}, \text{ if } E \text{ is a triangle,} \\ \dfrac{p+1}{2}, \text{ if } E \text{ is a quadrilateral.} \end{cases} \tag{12}$$

3.4 Optimal penalization parameters

Thanks to the previous results on the discrete linearized problem Problem $(\tilde{\mathcal{W}}_h)$, one can now construct a method to set automatically penalization parameters. They must be chosen to ensure the coercivity and continuity of the linearized discrete problem, i.e. $C^*(\epsilon) > 0$ and $\tilde{C}(\epsilon) > 0$. Moreover, using Céa's lemma 3.8, they are set to minimize the distance between the weak and discrete solutions.

Lemma 3.8 (Céa's lemma)**.** *Let V be a real Hilbert space with the norm $\|\cdot\|$. Let $a : V \times V \to \mathbb{R}$ be a bilinear form and $l : V \to \mathbb{R}$ a linear form satisfying the Lax-Milgram theorem. Let V_h be a closed subspace of V. Then there exists a unique $u_h \in V_h$ such that*

$$\forall\ v_h \in V_h,\ a(u_h, v_h) = l(v_h)\ \textit{and}\ \|u - u_h\| \leq \frac{\tilde{C}}{C^*}\|u - v\|, \quad \forall v \in V_h$$

where $\tilde{C}$ is the continuity constant and C^* the coercivity constant.

Firstly as a reminder, positivity of continuity and coercivity constants enforce that for all $n \in \{0, ..., N-1\}, \varepsilon^{(n)} < 2$ and $\forall n \in \{0, ..., N\},\ \sigma_n > \sigma_n^*$. They are given by :

$$C^*(\epsilon) = \min\Big\{ \min_{n=0,..,N-1} \Big(K_0^{(n)}(1 - \frac{\varepsilon^{(n)}}{2})\Big), \sigma_0 - \sigma_0^*, \sigma_N - \sigma_N^*, \min_{n=1,..,N-1} \Big(\frac{\sigma_n - \sigma_n^*}{2}\Big)\Big\},$$

and

$$\tilde{C}(\epsilon) = \max_{n=0,...,N-1} \left(K_1^{(n)}\right) + \sqrt{\max_{n=0,...,N-1} \left(2\varepsilon^{(n)} K_1^{(n)}\right)} \sqrt{\max\left(\sigma_0^*, \sigma_N^*, \max_{n=1,...,N-1}\left(\frac{\sigma_n^*}{2}\right)\right)}$$
$$+ \max\left(\sigma_0, \sigma_N, \max_{n=1,...,N-1}\left(\frac{\sigma_n}{2}\right)\right).$$

For the sake of simplicity, let us consider that the variable ε is the same for every element: $\forall n \in \{0, ..., N-1\}, \varepsilon^{(n)} = \varepsilon < 2$, and in addition, because penalization parameters are bounded below, let us consider that they are above the lower bound of an amount α constant for every element:

$$\forall \alpha > 1, \forall n \in \{1, ..., N-1\},\ \sigma_n = \frac{\alpha}{2\varepsilon}\tilde{\sigma}_n^*,\ \sigma_0 = \frac{\alpha}{\varepsilon}\tilde{\sigma}_0^*,\ \sigma_N = \frac{\alpha}{\varepsilon}\tilde{\sigma}_N^* \text{ with } \tilde{\sigma}_n^* = \frac{(K_1^{(n)} C_{\mathrm{tr},p-1}^{(n)})^2}{K_0^{(n)}}.$$

Using previous assumptions it can be noticed that C^* and $\tilde{C}$ are functions of ε and α and can be rewritten:

$$C^*(\alpha, \varepsilon) = \min\Big\{K_0(1 - \frac{\varepsilon}{2}), \frac{\alpha - 1}{\varepsilon}\tilde{\sigma}_0, \frac{\alpha - 1}{\varepsilon}\tilde{\sigma}_N, \min_{n=1,..,N-1}\Big(\frac{\alpha - 1}{\varepsilon}\frac{\tilde{\sigma}_n}{4}\Big)\Big\}$$

and

$$\tilde{C}(\epsilon) = K_1 + \sqrt{2\varepsilon K_1}\sqrt{\frac{1}{\varepsilon}\max\left(\tilde{\sigma}_0^*, \tilde{\sigma}_N^*, \max_{n=1,...,N-1}\left(\frac{\tilde{\sigma}_n^*}{4}\right)\right)} + \frac{\alpha}{\varepsilon}\max\left(\tilde{\sigma}_0, \tilde{\sigma}_N, \max_{n=1,...,N-1}\left(\frac{\tilde{\sigma}_n}{4}\right)\right)\ .$$

One can see that two quantities are involved in the two previous definitions:

$$\tilde{\sigma}_{min} = \min\Big\{\tilde{\sigma}_0, \tilde{\sigma}_N, \min_{n=1,..,N-1}\Big(\frac{\tilde{\sigma}_n}{4}\Big)\Big\} \text{ and } \tilde{\sigma}_{max} = \max\Big\{\tilde{\sigma}_0, \tilde{\sigma}_N, \max_{n=1,..,N-1}\Big(\frac{\tilde{\sigma}_n}{4}\Big)\Big\}$$

to have the final write:

$$C^*(\alpha, \varepsilon) = \min\Big\{K_0(1 - \frac{\varepsilon}{2}), \frac{\alpha - 1}{\varepsilon}\tilde{\sigma}_{min}\Big\} \text{ and } \tilde{C}(\alpha, \varepsilon) = K_1 + \sqrt{2K_1\tilde{\sigma}_{max}} + \frac{\alpha}{\varepsilon}\tilde{\sigma}_{max}$$

These new expressions of C^* and $\tilde{C}$ show that C^* has two different states and $\tilde{C}$ is continuous concerning α and ε. The aim of this section can now be reformulated as find α and ε such that $\gamma(\alpha, \varepsilon) = \dfrac{\tilde{C}(\alpha, \varepsilon)}{C^*(\alpha, \varepsilon)}$

is minimal. First, C^* and $\tilde{C}$ are studied separately, then γ is observed. C^* has two different states, is continuous and well defined for all $(\alpha,\varepsilon)\in(1,+\infty)\times(0,2)$. It can be rewritten as follows:

$$\forall(\alpha,\varepsilon)\in(1,+\infty)\times(0,2),$$
$$C^*(\alpha,\varepsilon)=\begin{cases}\dfrac{\alpha-1}{\varepsilon}\tilde{\sigma}_{min}, & \text{if } \alpha\leq\alpha^*(\epsilon)\\ K_0(1-\dfrac{\varepsilon}{2}), & \text{otherwise}\end{cases} \text{ with } \alpha^*(\epsilon)=\frac{K_0}{2\tilde{\sigma}_{min}}\varepsilon(2-\varepsilon)+1\ .$$

$\tilde{C}$ is continuous and well defined for all $(\alpha,\varepsilon)\in(1,+\infty)\times(0,2)$. C^* and $\tilde{C}$ are now explicitly characterized and now $\gamma(\alpha,\varepsilon)=\dfrac{\tilde{C}(\alpha,\varepsilon)}{C^*(\alpha,\varepsilon)}$ can be studied. $(\alpha_{opt},\varepsilon_{opt})$ are looked for such that γ is minimal and it is given by:

$$\forall(\alpha,\varepsilon)\in(1,+\infty)\times(0,2),$$
$$\gamma(\alpha,\varepsilon)=\begin{cases}\dfrac{\varepsilon}{\tilde{\sigma}_{min}(\alpha-1)}\left(K_1+\sqrt{2K_1\tilde{\sigma}_{max}}\right)+\dfrac{\alpha}{\alpha-1}\dfrac{\tilde{\sigma}_{max}}{\tilde{\sigma}_{min}} & \text{, if } \alpha\leq\alpha^*(\epsilon)\\ \dfrac{2}{K_0(2-\varepsilon)}\left(K_1+\sqrt{2K_1\tilde{\sigma}_{max}}+\dfrac{\alpha}{\varepsilon}\tilde{\sigma}_{max}\right) & \text{, otherwise}\end{cases}$$

γ is studied on its different open subdomains and the boundary between them. On $\mathcal{D}_1$, for all $(\alpha,\varepsilon)\in(\alpha^*(\varepsilon),+\infty)\times(0,2)$ it gives:

$$\gamma(\alpha,\varepsilon)=a\frac{1}{2-\varepsilon}+b\frac{\alpha}{\varepsilon(2-\varepsilon)} \text{ with } a=2\frac{K_1+\sqrt{2K_1\tilde{\sigma}_{max}}}{K_0} \text{ and } b=2\frac{\tilde{\sigma}_{max}}{K_0}.$$

Then, looking at its variations, it gives that:

$$\partial_\varepsilon\gamma(\alpha,\varepsilon)\begin{cases}<0 & \text{, if } 0<\varepsilon<\varepsilon^*\\ =0 & \text{, if } \varepsilon=\varepsilon^*\\ >0 & \text{, if } \varepsilon^*<\varepsilon<2\end{cases} \text{ and } \partial_\alpha\gamma(\alpha,\varepsilon)=\frac{b}{\varepsilon(2-\varepsilon)}>0$$

with $\varepsilon^*=\frac{\sqrt{b(2a+b)}-b}{a}>0$. And finally noting that $\gamma\to+\infty$ when $\alpha\to+\infty$ and when $\varepsilon\to0$ or $\varepsilon\to2$ it gives that γ is minimal for $\varepsilon=\varepsilon^*$ and $\alpha\to\alpha^*(\varepsilon^*)$.

On $\mathcal{D}_2$, for all $(\alpha,\varepsilon)\in(1,\alpha^*(\varepsilon))\times(0,2)$ it gives:

$$\gamma(\alpha,\varepsilon)=\frac{\varepsilon}{\tilde{\sigma}_{min}(\alpha-1)}\left(K_1+\sqrt{2K_1\tilde{\sigma}_{max}}\right)+\frac{\alpha}{\alpha-1}\frac{\tilde{\sigma}_{max}}{\tilde{\sigma}_{min}}.$$

Then, looking at its variations, it gives that:

$$\partial_\varepsilon\gamma(\alpha,\varepsilon)=\frac{K_1+\sqrt{2K_1\tilde{\sigma}_{max}}}{\tilde{\sigma}_{min}(\alpha-1)}>0 \text{and } \partial_\alpha\gamma(\alpha,\varepsilon)=-\frac{1}{(\alpha-1)^2}\left(\frac{\varepsilon}{\tilde{\sigma}_{min}}\left(K_1+\sqrt{2K_1\tilde{\sigma}_{max}}\right)+\frac{\tilde{\sigma}_{max}}{\tilde{\sigma}_{min}}\right)<0.$$

And finally noting that $\gamma\to+\infty$ when $\alpha\to1$ it gives that γ is minimal for $\alpha\to\alpha^*(\varepsilon)$. On the boundary between $\mathcal{D}_1$ and $\mathcal{D}_2$, for all $\alpha=\alpha^*(\varepsilon)$ and $\varepsilon\in(0,2)$ it gives:

$$\gamma(\alpha^*(\varepsilon),\varepsilon)=a\frac{1}{2-\varepsilon}+b\frac{1}{\varepsilon(2-\varepsilon)}+c \text{ with } a=2\frac{K_1+\sqrt{2K_1\tilde{\sigma}_{max}}}{K_0},\ b=2\frac{\tilde{\sigma}_{max}}{K_0} \text{ and } c=\frac{\tilde{\sigma}_{max}}{\tilde{\sigma}_{min}}.$$

Then, looking at its variations, it gives that:

$$\partial_\varepsilon\gamma(\alpha^*(\varepsilon),\varepsilon)\begin{cases}<0 & \text{, if } 0<\varepsilon<\varepsilon_{opt}\\ =0 & \text{, if } \varepsilon=\varepsilon_{opt}\\ >0 & \text{, if } \varepsilon_{opt}<\varepsilon<2\end{cases} \text{ with } \varepsilon_{opt}=\frac{\sqrt{b(2a+b)}-b}{a}>0.$$

The expression of $(\alpha_{opt}, \varepsilon_{opt})$ can be summarized as follows:

$$\varepsilon_{opt} = \frac{\sqrt{b(2a+b)} - b}{a} \text{ with } a = 2\frac{K_1 + \sqrt{2K_1\tilde{\sigma}_{max}}}{K_0} \text{ and } b = 2\frac{\tilde{\sigma}_{max}}{K_0} \text{ and } \alpha_{opt} = \frac{K_0}{2\tilde{\sigma}_{min}}\varepsilon_{opt}(2-\varepsilon_{opt}) + 1.$$

Finally in one dimension, the auto-calibration of penalization parameters is given by:

$$\forall n \in \{1, ..., N-1\},\ \sigma_n = \frac{\alpha_{opt}}{2\varepsilon_{opt}}\tilde{\sigma}_n^*,\ \sigma_0 = \frac{\alpha_{opt}}{\varepsilon_{opt}}\tilde{\sigma}_0^*,\ \sigma_N = \frac{\alpha_{opt}}{\varepsilon_{opt}}\tilde{\sigma}_N^* \text{ with } \tilde{\sigma}_n^* = \frac{(K_1^{(n)}C_{\mathrm{tr},p-1}^{(n)})^2}{K_0^{(n)}}.$$

In two dimensions, the auto-calibration of penalization parameters is given by:

$$\begin{cases} \forall F \in \mathcal{F}^{\mathrm{in}},\ \sigma_E^{in} = \frac{\alpha_{opt}}{2\varepsilon_{opt}}\sigma_E^*,\ \sigma_{E_r}^{in} = \frac{\alpha_{opt}}{2\varepsilon_{opt}}\sigma_{E_r}^* \\ \forall F \in \mathcal{F}^{\partial},\ \sigma_E^{\partial} = \frac{\alpha_{opt}}{\varepsilon_{opt}}\sigma_E^* \end{cases} \text{ with } \sigma_E^* = \frac{D^E\left(K_1^E C_{\mathrm{tr},p-1}^E\right)^2}{2\varepsilon^E K_0^E}$$

and D^E the number of edges of the element E and $C_{\mathrm{tr},p-1}^E$ the trace constant defined in Equation (12).

3.5 Convergence of the discrete linearized weak problem to the continuous linearized weak problem

Previously, it has been proven that the Problem $(\tilde{\mathcal{W}}_h)$ has a unique solution. This problem is part of a fixed point method, and it has been proven in Section 3.2 that this fixed point has a unique solution also. To solve the non-linear weak formulation Problem $(\mathcal{W})$, one step needs to be added to prove the well-posedness of the problem. It is addressed in the following; the goal is to prove that the solution of Problem $(\tilde{\mathcal{W}}_h)$ converges towards the solution of Problem $(\tilde{\mathcal{W}})$ and prove that the bilinear form $\tilde{a}_h$ of Problem $(\tilde{\mathcal{W}}_h)$ converges to Problem $(\mathcal{W})$.

The work in this section is based on the book of Di Pietro and Ern published in 2011 [31]. They proved convergence in the case of a Symmetric Interior Penalty Galerkin method and sketch the proof in the case of an Incomplete Interior Penalty method. The following study provides detailed proof of the IIPG case.

The key idea is to revisit the concept of consistency and introduce a new point of view based on asymptotic consistency. This new form of consistency and the usual stability of the discrete bilinear form are the two main ingredients for asserting convergence to the minimal regularity solutions. The discrete bilinear form a_h needs to be reformulated to consider only the contribution of K on the mesh elements, not the interfaces; consequently, lifting operators are introduced. They map functions defined on mesh faces to functions defined on mesh elements. In the context of DG methods, liftings act on interfaces and boundary jumps. Bassi, Rebay *et al.* introduced them [2] in the context of compressible flows and analyzed by Brezzi, Manzini *et al.* [5] in the context of the Poisson problem. Liftings have many useful applications. They can be combined with the gradient to define discrete gradients. Discrete gradients play an essential role in the design and analysis of DG methods. Indeed, they can be used to formulate the discrete problem locally on each element using numerical fluxes.

Liftings: Definition

For any point x_n, and for all $\varphi \in L^2(\{x_n\})$ the lifting operator $r_n^p : L^2(\{x_n\}) \longrightarrow \mathcal{V}_0^p(\mathcal{E}_h)$ is defined as the solution of the following problem:

$$\int_\Omega r_n^p(\varphi)\tau_h dx = \{\!|\tau_h|\!\}_{x_n}\varphi(x_n), \quad \forall \tau_h \in \mathcal{V}_0^p(\mathcal{E}_h).$$

For any v in $\mathcal{V}_0^p(\mathcal{E}_h)$, the global lifting of its interface and boundary jumps is defined as follows:

$$R_h^p([\![v]\!]) := \sum_{n=0}^{N} r_n^p([\![v]\!]) \in \mathcal{V}_0^p(\mathcal{E}_h).$$

Discrete gradients: Definition

The discrete gradient operator $G_h^p : \mathcal{V}_0^p(\mathcal{E}_h) \longrightarrow L^2(I_n)$ is defined as follows: for all v in $\mathcal{V}_0^p(\mathcal{E}_h)$,

$$G_h^p(v) := \nabla_h v - R_h^p([\![v]\!]). \tag{13}$$

In addition, there exists a bound on the discrete gradient operator:

$$\|G_h^p(v)\|_{L^2(\Omega)} \leq \alpha \|v\| \tag{14}$$

where $\|\cdot\|$ is the norm associated with the IIPG formulation defined Equation (10).

Theorem 3.3 (Regularity of the limit and weak asymptotic consistency of discrete gradients)**.** *Let $p \geq 0$. Let v_h be a sequence in $\mathcal{V}_0^p(\mathcal{E}_h)$ bounded by the $\|.\|$-norm. Then, there is a function $v \in H_0^1(\Omega)$ such that as $h \to 0$, up to a subsequence,*

$$v_h \to v \text{ strongly in } L^2(\Omega),$$

and for all $p \geq 0$, the discrete gradients defined by Equation (13) *are such that*

$$G_h^p(v_h) \rightharpoonup v' \text{ weakly in } L^2(\Omega).$$

Proof of Theorem 3.3 is available in [31, pp. 194-195].
Because of the shape of the IIPG formulation, the modified discrete gradient operator $\hat{G}_h^p : \mathcal{V}_0^p(\mathcal{E}_h) \longrightarrow L^2(I_n)$ is defined as follows: for all v in $\mathcal{V}_0^p(\mathcal{E}_h)$,

$$\hat{G}_h^p(v) := \nabla_h v.$$

Using liftings and discrete gradients, surface contributions of the flux in Equation (11) are transformed to volume contribution. It makes working with the bilinear form $\tilde{a}_h$ easier. For a given $\bar{u} \in \mathcal{V}_0^p(\mathcal{E}_h)$ it can be rewritten as follows:

$$\begin{aligned}
&\forall u_h, v_h \in \mathcal{V}_0^p(\mathcal{E}_h),\\
&\tilde{a}_h(u_h, v_h; \bar{u}) = \sum_{n=0}^{N-1} \int_{I_n} K(x,\bar{u}) \nabla_h u_h \nabla_h v_h dx - \sum_{n=0}^{N} \{\!\{ K(x,\bar{u}) \nabla_h u_h \}\!\}_{x_n} [\![v_h]\!]_{x_n} + s_h(u_h, v_h)\\
&= \sum_{n=0}^{N-1} \int_{I_n} K(x,\bar{u}) \hat{G}_h^p(u_h) \nabla_h v_h dx - \sum_{n=0}^{N} \sum_{m=0}^{N-1} \int_{I_m} K(x,\bar{u}) r_n^p([\![v_h]\!]) \hat{G}_h^p(u_h) + s_h(u_h, v_h)\\
&= \sum_{n=0}^{N-1} \int_{I_n} K(x,\bar{u}) \hat{G}_h^p(u_h) \nabla_h v_h dx - \sum_{n=0}^{N-1} \int_{I_n} K(x,\bar{u}) R_h^p([\![v_h]\!]) \hat{G}_h^p(u_h) + s_h(u_h, v_h)\\
&= \sum_{n=0}^{N-1} \int_{I_n} K(x,\bar{u}) \hat{G}_h^p(u_h) G_h^p(v_h) dx + s_h(u_h, v_h)
\end{aligned}$$

with

$$\forall u_h, v_h \in \mathcal{V}_0^p(\mathcal{E}_h),\ s_h(u_h, v_h) = \frac{\sigma_0}{h} [\![u_h]\!]_{x_0} [\![v_h]\!]_{x_0} + \sum_{n=1}^{N-1} \frac{\sigma_{n-1} + \sigma_n}{2h} [\![u_h]\!]_{x_n} [\![v_h]\!]_{x_n} + \frac{\sigma_N}{h} [\![u_h]\!]_{x_N} [\![v_h]\!]_{x_N}\ .$$

Consider that $(\sigma_n)_{n=0,\ldots,N}$ are chosen according to the Lemma 3.3 that implies discrete coercivity in the $\|.\|$-norm, and hence well-posedness of the discrete linearized problem $(\tilde{\mathcal{V}}_h)$.

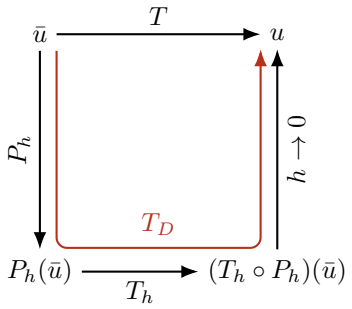


Figure 6: Scheme of the whole loop of resolution with the different linearization methods

Definition 3.1 (Asymptotic adjoint consistency)**.** *The discrete bilinear form a_h is asymptotically adjoint consistent with the exact bilinear form a on $\mathcal{V}_0^p(\mathcal{E}_h)$ if for any subsequence v_h in $\mathcal{V}_0^p(\mathcal{E}_h)$ bounded in the $\|.\|$-norm and for any smooth function $\varphi \in C_0^\infty(\Omega)$, there is a subsequence φ_h in $\mathcal{V}_0^p(\mathcal{E}_h)$ converging to φ in the $\|.\|$-norm and such that, up to a subsequence*

$$\lim_{h\to 0} a_h(v_h, \varphi_h) = a(v, \varphi) = \int_\Omega v'\varphi' dx$$

where $v \in H_0^1(\Omega)$ is the limit of the subsequence identified in Theorem 3.3.

Lemma 3.9 (Asymptotic adjoint consistency of $\tilde{a}_h$)**.** *The discrete bilinear form $\tilde{a}_h$ of Problem $(\tilde{\mathcal{W}}_h)$ is asymptotically adjoint consistent with the exact bilinear form $\tilde{a}$ of Problem $(\tilde{\mathcal{W}})$ on $\mathcal{V}_0^p(\mathcal{E}_h)$.*

Finally, we deduce the following result:

Theorem 3.4 (Convergence to minimal regularity solutions)**.** *Let $p \geq 1$. Let u_h be a sequence of approximate solutions generated by solving the discrete linearized problem $(\tilde{\mathcal{V}}_h)$ with $\tilde{a}_h$ defined by Equation (11) and with penalty parameters ensuring coercivity. Then as $h \to 0$*

$$\begin{aligned} &u_h \longrightarrow u \text{ strongly in } L^2(\Omega) \\ &\nabla_h u_h \longrightarrow u' \text{ strongly in } L^2(\Omega) \\ &|u_h|_J \to 0 \end{aligned}$$

where $u \in H_0^1(\Omega)$ is the unique solution of the strong problem.

Proofs of Lemma 3.9 and Theorem 3.4 can be found in Appendix A .

3.6 Concluding results

In the current section, several theorems have been proven. It is proven that there exists a unique solution to Problem $(\mathcal{W})$ using Lemma 3.1 and Lemma 3.2 . Then it is proven that for a given $\bar{u}$, there exists a unique solution to Problem $(\tilde{\mathcal{W}}_h)$ using Lemma 3.1 . Lastly it is proven that for a given $\bar{u}$, the solution of Problem $(\tilde{\mathcal{W}}_h)$ converges to the solution of Problem $(\tilde{\mathcal{W}})$. These results proven in a general case for a given $\bar{u}$ can be used to solve the toy problem. Figure 6 gives a graphical representation of the whole loop of resolution with different paths.

The non-linear problem, Problem $(\mathcal{W})$ can be linearized directly at the continuous level by employing a fixed point method. The continuous level linearization

$$\begin{aligned} T : H_0^1(\Omega) &\longrightarrow H_0^1(\Omega) \\ \bar{u} &\longmapsto T(\bar{u}) = u \end{aligned}$$

stands for : find u solution of Problem $(\tilde{\mathcal{W}})$ for a given $\bar{u} \in H_0^1(\Omega)$. One can define the following sequence defined by $u^0 \in H_0^1(\Omega)$ an initial guess and $u^{n+1} = T(u^n)$ for $n \in \mathbb{N}$. Lemma 3.1 and Lemma 3.2 ensures that taking $\lim_{n\to\infty} u^n$ gives the solution of Problem $(\mathcal{W})$.

A discretization step is needed to compute the solution of Problem $(\mathcal{W})$. Consequently, the projector $P_h : H_0^1(\Omega) \longrightarrow \mathcal{V}_0^p(\mathcal{E}_h)$ is introduced. It projects a function living in an infinite-dimensional space to a finite-dimensional space, especially it projects a function to the DG space $V_0^p(\mathcal{E}_h)$. Then at a discrete level the linearization method

$$\begin{aligned} T_h : V_0^p(\mathcal{E}_h) &\longrightarrow V_0^p(\mathcal{E}_h) \\ \bar{u} &\longmapsto T_h(\bar{u}) = u_h \end{aligned}$$

stands for : find u_h discrete solution of Problem $(\tilde{\mathcal{W}}_h)$ for a given $\bar{u} \in V_0^p(\mathcal{E}_h)$. One can notice that for a given $\bar{u} \in H_0^1(\Omega)$, it has been proven (Theroem 3.4) that $(T_h \circ P_h)(\bar{u}) = u_h$ converges to u given by $T_C(\bar{u})$.

Lastly the linearization method of Problem $(\mathcal{W})$ going through a discretization step is defined as

$$\begin{aligned} T_D : H_0^1(\Omega) &\longrightarrow H_0^1(\Omega) \\ \bar{u} &\longmapsto T_D(\bar{u}) = \lim_{h\to 0}(T_h \circ P_h)(\bar{u}) = u \end{aligned}$$

Using T_D one can define a new sequence $v^0 \in H_0^1(\Omega)$ an initial guess and $v^{n+1} = T_D(v^n) = \lim_{h\to 0}(T_h \circ P_h)(v^n)$ for $n \in \mathbb{N}$. Taking the limit when n goes to infinity gives the solution of Problem $(\mathcal{W})$.

The previously explained method use two limits, h goes to 0 then n goes to infinity. One can also consider limits in the opposite order. Using proof of Lemma 3.1 applied to the non-linear discrete problem and then using Theroem 3.4 one can prove that the solution of the non-linear discrete problem converges to the solution of the non-linear continuous problem.

4 Numerical results

Following the numerical methods and theoretical results presented in the previous sections, the `RIVAGE` code is validated against numerical test cases. Two analytical test cases are used to compute convergence rates and validate the code. These analytical test cases are obtained by considering the problem's aimed solution and choosing the source term according to the solution and the hydraulic conductivity function. They are built upon the non-linear Poisson's equation. The first case is a non-linear one-dimensional problem in its stationary form. The second case is a non-linear two-dimensional problem in its stationary form. These numerical experiments are inspired by literature. In 2008, Rivière [34] and in 2021, Clément [9] computed convergence rates for linear problems also for non-linear problems.

Stationary problems are considered since theoretical results are given on this type of problem. Moreover, they are more difficult to solve since they solve the problem at infinite time. Consequently, the non-linear solver has to find the solution without getting time sub-steps

Experimental test cases are solved with the `RIVAGE` code. These problems aim at confirming the performance of the adaptive strategy proposed in this work. Moreover, they allow to test `RIVAGE` of problems encountered in the hydrology field. These experiments are based on the work of Haverkamp [22] and Vauclin [43].

4.1 One-dimensional analytical test case

For this first test case, theoretical convergence rates of the IIPG methods are checked, and numerical stability is evaluated concerning penalty values and penalization methods. The following problem is considered:

$$\begin{cases} -\partial_x(K(u)\partial_x u) &= f(x) \text{ in } \Omega = [-1, 1] \\ u(-1) = 1, \\ u(1) = -1, \end{cases}$$

		p = 1			**p = 2**			**p = 3**		
σ	$\mathbf{N_x}$	$\mathbf{L^2}$**-error**	**r**	**t(s)**	$\mathbf{L^2}$**-error**	**r**	**t(s)**	$\mathbf{L^2}$**-error**	**r**	**t(s)**
1	20	$3.21\,10^{-1}$		0.21	$1.33\,10^{-1}$		1.94	$2.29\,10^{-4}$		6.21
-	40	$1.29\,10^{-1}$	1.31	0.46	$3.41\,10^{-2}$	1.97	3.17	$1.42\,10^{-5}$	4.01	11.85
-	80	$3.77\,10^{-2}$	1.78	1.02	$8.53\,10^{-3}$	2.00	5.97	$8.88\,10^{-7}$	4.00	23.94
-	160	$9.83\,10^{-3}$	1.94	2.08	$2.13\,10^{-3}$	2.00	12.09	$5.62\,10^{-8}$	3.98	52.83
-	Fitted	—	1.69	—	—	1.99	—	—	4.00	
100	20	$8.33\,10^{-3}$		0.21	$1.36\,10^{-3}$		1.48	$2.33\,10^{-6}$		5.89
-	40	$2.10\,10^{-3}$	1.99	0.51	$3.41\,10^{-4}$	2.00	2.97	$1.44\,10^{-7}$	4.02	11.87
-	80	$5.27\,10^{-4}$	2.00	1.03	$8.53\,10^{-5}$	2.00	5.94	$9.74\,10^{-9}$	3.89	24.03
-	160	$1.31\,10^{-4}$	2.00	2.08	$2.13\,10^{-5}$	2.00	12.16	$1.37\,10^{-9}$	2.83	53.20
-	Fitted	—	1.99	—	—	2.00	—	—	3.61	
auto	20	$3.53\,10^{-2}$		0.24	$1.69\,10^{-2}$		1.43	$3.46\,10^{-6}$		5.88
-	40	$8.88\,10^{-3}$	1.99	0.51	$4.40\,10^{-3}$	1.94	2.86	$1.61\,10^{-7}$	4.42	11.92
-	80	$2.17\,10^{-3}$	2.03	1.05	$1.13\,10^{-3}$	1.95	5.95	$9.45\,10^{-9}$	4.10	24.07
-	160	$5.32\,10^{-4}$	2.03	2.55	$2.90\,10^{-4}$	1.97	12.13	$1.33\,10^{-9}$	2.82	53.25
-	Fitted	—	2.02	—	—	1.96	—	—	3.81	

Table 3: L^2-error, convergence rates and number of iterations for the one-dimensional benchmark.

with $K(u) = \tanh(5u) + 1.01$ and f obtained by replacing u by u_{ex} in the problem. The chosen analytical solution is $u_{ex}(x) = -\sin(\frac{\pi}{2}x)$. The analytical solution is chosen not to be polynomial but to span the interval $[-1, 1]$. The hydraulic conductivity is chosen to have a non-linear problem with a similar shape of law given in Table 1. tanh has been chosen because it is a smooth function convenient for the computation of convergence rates and looks like constitutive laws for RE. Moreover, a factor 200 between the maximum and the minimum value of K with $K_0 = 0.01$. The problem is solved with the IIPG method. Three types of penalization are used. The first one $\sigma^E = \sigma = 1$ for all $E \in \mathcal{E}$, the second one $\sigma^E = \sigma = 100$ for all $E \in \mathcal{E}$ and the third one σ^E are auto-calibrated using the method presented in Section 3. For each type of penalization, the solution is approximated by a piecewise linear function ($p = 1$), a piecewise quadratic function ($p = 2$), and a piecewise cubic function ($p = 3$). Moreover, lastly, four different mesh sizes are used $N_x = 20, 40, 80, 160$ with N_x the number of elements in the equally spaced partition of Ω.

Table 3 shows L^2-error and convergence rate for each computation. It can be noticed that computed convergence rates correspond to the theoretical ones found in literature [34] and [14, pp. 64-84]. For the IIPG formulation with penalization, p is odd is order $p + 1$, it is optimal, and if p is even the order is p, it is suboptimal. Moreover, for a penalization speed set by the user to 1 (outside of the range specified by theoretical results), errors are about 100 times greater than other computations. The fixed point method converges to a less accurate solution. Computation times are also given. It can be noticed that auto penalization is not greatly slower than user-defined penalization and can even be faster due to the quickest convergence of the iterative method.

Moreover, Figure 7 shows penalization values in the case of auto-calibration. One can observe that penalization values are not constant on Ω and vary according to the polynomial degree of approximation. On the domain, some part needs a small amount of penalization, whereas some others need a higher amount.

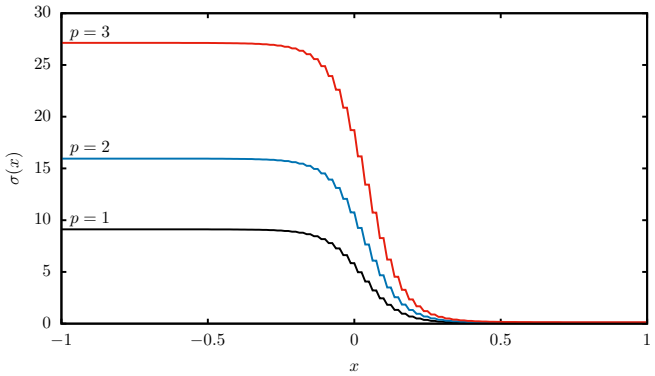


Figure 7: Penalization parameters for the one-dimensional test case in the case of auto penalization.

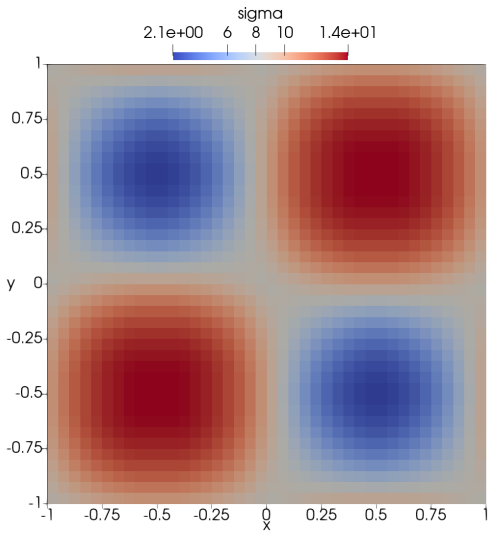


Figure 8: Penalization parameters for the two-dimensional test case in the case of auto penalization.

4.2 Two-dimensional analytical test case

This second experiments focuses on the ability of the IIPG method to solve RE in two dimensions. Its convergence rates are computed, and numerical stability is evaluated concerning penalty values and penalization methods. The following problem is considered:

$$\begin{cases} -\nabla \cdot (K(u)\nabla u) &= f(\mathbf{x}) \text{ in } \Omega = [-1,1] \times [-1,1] \\ u = 0 \text{ on } \partial\Omega, \end{cases}$$

with $K(u) = \tanh(u) + 1.01$ and similarly to the previous test case f is obtained by replacing u by u_{ex} in the problem. The chosen analytical solution is $u_{ex}(x,y) = \sin(\frac{\pi}{2}x)\sin(\frac{\pi}{2}y)$. The problem is solved similarly to the one-dimensional test case. Three types of penalization are used. The first one $\sigma^E = \sigma = 1$ for all $E \in \mathcal{E}$, the second one $\sigma^E = \sigma = 100$ for all $E \in \mathcal{E}$ and the third one σ^E are auto-calibrated. For each type of penalization, the solution is approximated by a piecewise linear function ($p = 1$), a piecewise quadratic function ($p = 2$), and a piecewise cubic function ($p = 3$). Lastly, three different meshes are used. They are all composed of quadrilaterals of identical size, and each space direction is discretized with $N = 10, 20, 40$ elements. It gives a mesh with $N_E = 100, 400, 1600$ elements. Table 4 shows L^2-error and convergence rate

		p = 1			**p = 2**			**p = 3**		
σ	$\mathbf{N_x}$	$\mathbf{L^2}$**-error**	**r**	**t(s)**	$\mathbf{L^2}$**-error**	**r**	**t(s)**	$\mathbf{L^2}$**-error**	**r**	**t(s)**
1	10	$6.45\,10^{-2}$		0.29	$4.83\,10^{-2}$		1.54	$8.60\,10^{-4}$		5.07
-	20	$1.51\,10^{-2}$	1.99	1.00	$1.11\,10^{-2}$	2.11	7.21	$4.69\,10^{-5}$	4.20	27.21
-	40	$3.53\,10^{-3}$	2.10	7.57	$2.65\,10^{-3}$	2.07	59.51	$2.74\,10^{-6}$	4.09	279.59
-	Fitted	—	2.10	—	—	2.09	—	—	4.15	
100	10	$3.80\,10^{-2}$		0.25	$2.02\,10^{-3}$		1.14	$7.32\,10^{-5}$		4.83
-	20	$9.53\,10^{-3}$	1.99	0.99	$2.72\,10^{-4}$	2.90	6.78	$4.59\,10^{-6}$	4.00	30.23
-	40	$2.38\,10^{-3}$	2.00	8.37	$4.08\,10^{-5}$	2.74	61.62	$2.87\,10^{-7}$	4.00	290.86
-	Fitted	—	2.00	—	—	2.82	—	—	4.00	
auto	10	$3.37\,10^{-2}$		0.25	$2.52\,10^{-3}$		1.15	$7.41\,10^{-5}$		4.93
-	20	$8.11\,10^{-3}$	2.06	1.03	$5.90\,10^{-4}$	2.09	6.88	$4.71\,10^{-6}$	3.98	30.15
-	40	$2.02\,10^{-3}$	2.00	8.49	$1.51\,10^{-4}$	1.96	60.85	$2.97\,10^{-7}$	3.99	288.50
-	Fitted	—	2.03	—	—	2.03	—	—	3.98	

Table 4: L^2-error, convergence rates and number of iterations for the two-dimensional benchmark.

for each computation. It can be noticed that computed convergence rates correspond to the theoretical ones found in literature [34] and [14, pp. 64-84]. For the IIPG formulation with penalization, p is odd in order $p+1$, it is optimal, and if p is even, the order is p and suboptimal. Moreover, for a penalization speed set by the user to 1 (outside of the range specified by theoretical results), errors are about 100 times greater than other computations. The fixed point method converges to a less accurate solution. Computation times are also given. It can be noticed that auto penalization is not greatly slower than user-defined penalization and can even be faster due to the quickest convergence of the iterative method as in the one-dimensional case.

Moreover, Figure 8 shows penalization values in the case of auto-calibration. One can observe that penalization values are not constant on Ω and vary according to the polynomial degree of approximation. On the domain, some part needs a small amount of penalization, whereas others need a higher amount.

4.3 Application to groundwater flows I: Haverkamp's test case

The two problems considered here, one-dimensional and two-dimensional, aim to validate the numerical resolution of RE using DG methods and auto-calibration of penalization parameters. Numerical results are compared to numerical simulations in the literature and experimental data.

The first experimental validation of solving RE with DG methods is a one-dimensional test case. The numerical results are compared with data sourced from the literature. This particular numerical test case was initially presented by Celia et al. [7]. It is based on an experiment conducted by Haverkamp et al. [22], who referred to the availability of a quasi-analytical solution provided by Philip [30]. Subsequently, it was used by others such as [37, 28], and represents a set of well-established test cases, for instance, see [29]. Despite its simplicity, this case offers insights into the fundamental physics of a wetting front resulting from infiltration.

This scenario involves the one-dimensional infiltration into a soil column measuring $40cm$ in height and $8cm$ in width. The hydraulic head at the top and bottom is governed by Dirichlet boundary conditions: $h_{top} = 19.3cm$ and $h_{bottom} = -61.5cm$, resulting in cumulative downward infiltration. The sides are impermeable. The initial condition is $h_0 = -61.5 + z\ cm$. Although this case is one-dimensional, it is solved on a two-dimensional domain. Therefore, homogeneous Neumann boundary conditions are applied along the boundary in the infiltration direction. For a visual representation of this setup, refer to Figure 9.

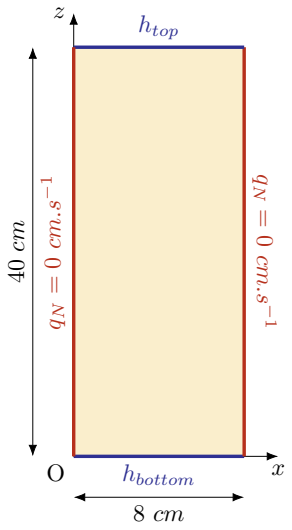


Figure 9: Haverkamp's test case configuration.

Hydraulic properties use Vachaud's relations in Table 1 with $A = 1.175\,10^6$, $B = 4.74$, $C = 1.611\,10^6, D = 3.96, Ks = 0.0094\ cm.s^{-1}, \theta_s = 0.287$ and $\theta_r = 0.075$. The simulation is done on a mesh of 160 elements along the z-axis. The solution is piecewise linear ($p = 1$), and time integration is BDF of order 2. Penalization parameters are set automatically using results from Section 3. In addition, stopping criteria are set to 10^{-6} for this computation. The solution to this problem is computed at $T = 600s$.

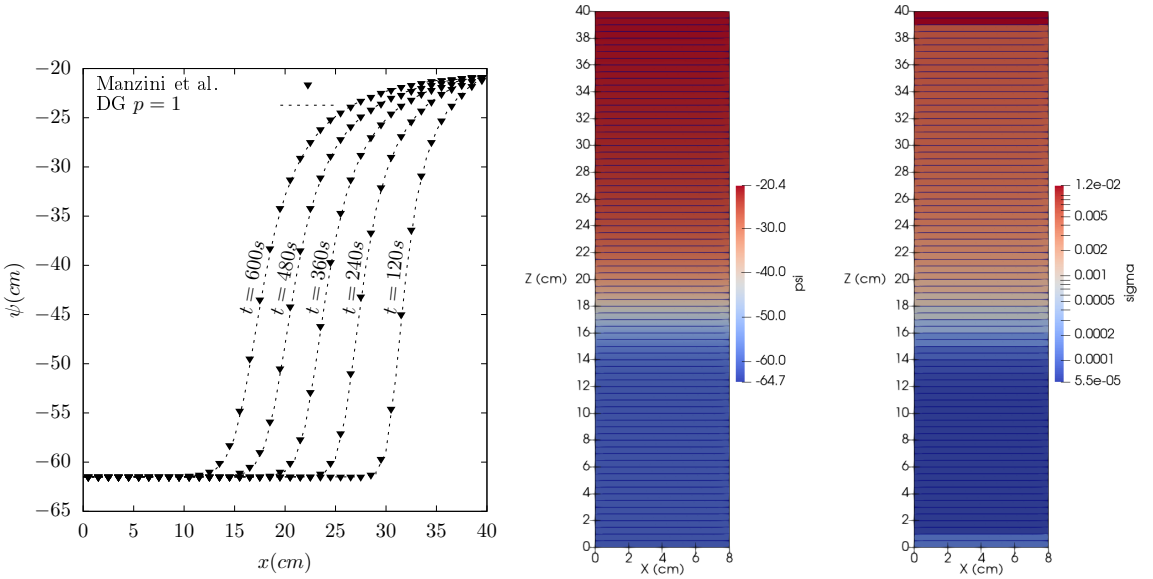
On Figure 10 are displayed the comparison of numerical results with results from Manzini et al. [28], the pressure head distribution at $t = 360s$ and the penalization parameters distribution at $t = 360s$. Numerical results are in good agreement with the literature results for this test case. The pressure head distribution shows a vertical progression of the wetting front with a steep transition from the initial ψ to ψ imposed at the boundary condition. Moreover, the distribution of penalization parameters shows that the penalization parameters are not constant on the whole domain and are higher on the wetting front.

This test case validates a real, evolving test case for the DG method. Moreover, it gives a good insight into the behavior of automatic penalization. Penalization parameters are auto-calibrated as long as the solution evolves. Moreover, automatic penalization impacts a full non-linear problem because the non-linear solver needs fewer iterations to converge to the solution.

4.4 Application to groundwater flows II: Vauclin's test case

Vauclin, Vachaud, and Khanji conducted a series of laboratory experiments in the 1970s, the details of which can be found in [43]. These experiments explored water table recharge and drainage in a slab of sandy soil. The work by Vauclin et al. [43] specifically focuses on simulating water flow recharge through a soil slab and provides experimental details and results. The experiment involved a 6 m by 2 m box, with only one half simulated due to symmetry. The left, top (for $x > 50\ cm$), and bottom sides were impervious, with a prescribed constant flux on the top for $x \leq 50\ cm$ of $\mathbf{u}_g \cdot \mathbf{n} = -14.8\ cm.h^{-1}$. The water level was maintained at a constant $h = 65\ cm$ in the ditch on the right for $z \leq 65\ cm$, while the remaining boundary on the right for $z > 65\ cm$ accounted for a seepage boundary condition. The initial state was at hydrostatic equilibrium with the water table at $z = 65\ m$. For further reference, please see Figure 11 for a schematic representation of the setup. The complete simulation of water table recharge by Vauclin et al. [43] has been used by numerous studies to evaluate their methods (see for instance, [13, 40, 45]). The MODFLOW code validation partially relies on this experimental dataset [39].

Hydraulic properties use Vachaud's relations in Table 1 with $A = 2.99\,10^6$, $B = 5.0$, $C = 40000, D =$



(a) Comparison with Manzini et al. [28]

(b) Pressure head distribution at $t = 360s$

(c) Penalization parameters distribution at $t = 360s$

Figure 10: Haverkamp's test case, numerical solution for 160 elements, $p = 1$ and BDF-2 method.

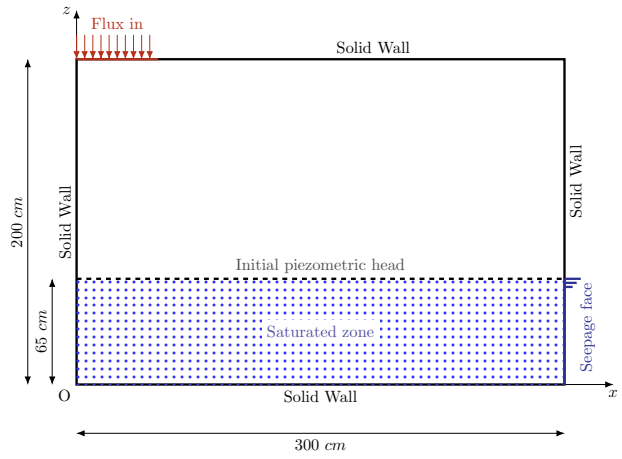


Figure 11: Vauclin's test case configuration.

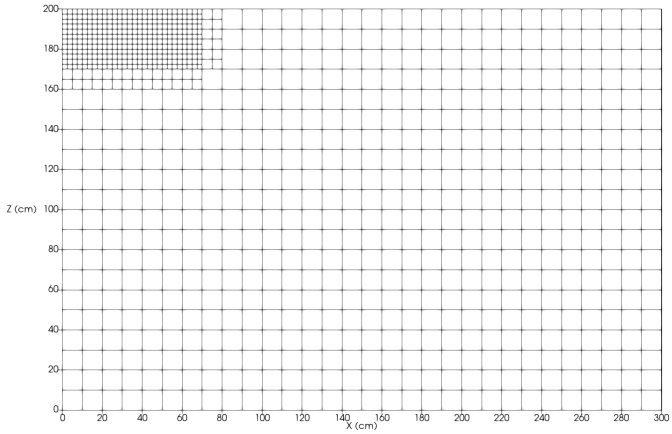


Figure 12: Vauclin's test case, initial mesh.

$2.9, Ks = 35\ cm.h^{-1}, \theta_s = 0.3$ and $\theta_r = 0.0$. The simulation is carried on an evolving mesh. The mesh is adapted along the computation according to the gradient of h. Mesh adaptive parameters are set to $\beta_c = 50$ and $\beta_r = 50$. The solution is sought piecewise linear ($p = 2$) and time integration is BDF of order 3. Penalization parameters are set automatically using results from Section 3. In addition, stopping criteria are set to 10^{-6} for this computation. The solution of this problem is computed until $T = 10h$.

In the initial mesh displayed in Figure 12, the refinement below the water entry edge aims to assist in simulating the steep wetting front. Figure 13 compares the water table's position at $t = 2, 3, 4, 8\ h$ with data from Vauclin et al. [43]. The numerical results closely match the experimental profile, although there are small discrepancies in the middle of the water table, which may be due to the non-perfect isotropic and homogeneous nature of the sandy soil.

Figures 14 and 15 illustrate the field distribution of hydraulic head, flux, and the positions of the water table and capillary fringe at $\theta = 0.29$. These figures also show the isolines of the hydraulic head. The numerical results are in agreement with the data from Vauclin et al. [43].

Additionally, in Figure 16, the evolution of penalization parameters during the computation is presented. At selected times, the evolution of the mesh reflects the capture of the steep front.

Finally, Figure 17 displays the evolution of time-steps and the number of elements over time. The adaptation of time-steps and the number of elements is evident, with the time-steps initially small due to the strong non-linearity induced by the steep wetting front. As the front smoothens, the number of elements decreases, stabilizing at $N_{ele} = 600$ after $t = 3\ h$.

This test case is a test case is a typical problem where auto-calibration of penalization parameters is essential. Since the problem is strongly non-linear and evolving, with a basic penalization and user defined parameters, the non-linear solver failed to capture the solution or necessitates some combination of fixed point solver and Newton-Raphson method such as in the work of [8].

A Proofs on theoretical results

Proof of Lemma 3.3. For a given $\bar{u} \in \mathcal{V}_0^p(\mathcal{E}_h)$ and choosing $v_h = u_h$ in (11) yields

$$\forall u_h \in \mathcal{V}_0^p(\mathcal{E}_h), \quad \tilde{a}_h(u_h, u_h) = \sum_{n=0}^{N-1} \int_{I_n} K(x, \bar{u})(u_h')^2 dx - \sum_{n=0}^{N} \{\!\{ K(x, \bar{u}) u_h' \}\!\}_{x_n} [\![u_h]\!]_{x_n} + \frac{\sigma_0}{h} [\![u_h]\!]_{x_0}^2 + \sum_{n=1}^{N-1} \frac{\sigma_{n-1} + \sigma_n}{2h} [\![u_h]\!]_{x_n}^2 + \frac{\sigma_N}{h} [\![u_h]\!]_{x_N}^2$$

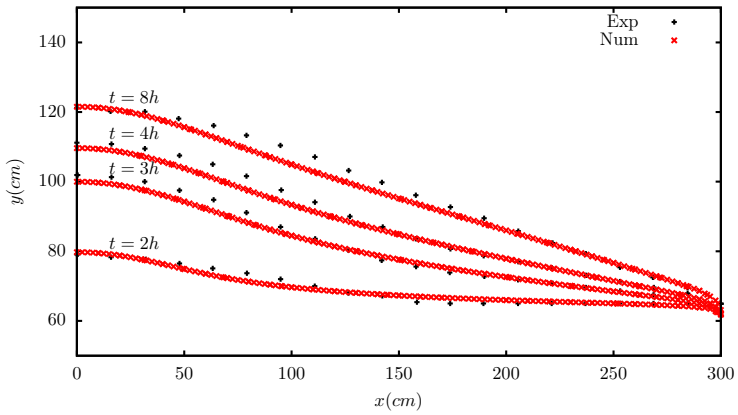
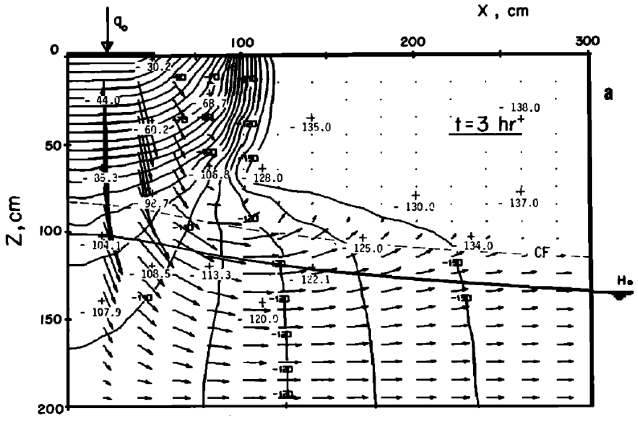
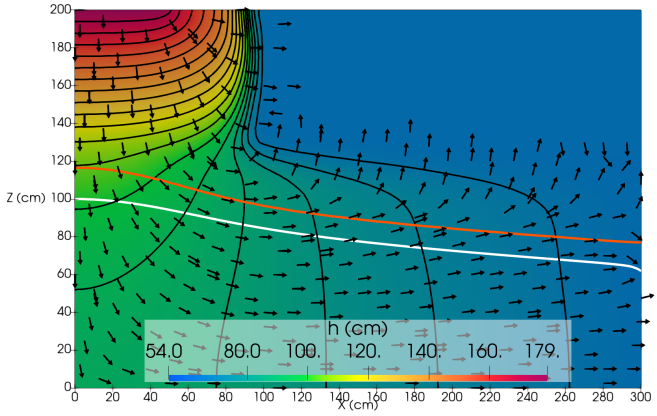


Figure 13: Vauclin's test case, numerical water table position compared to experimental data from Vaulin et al. [43].

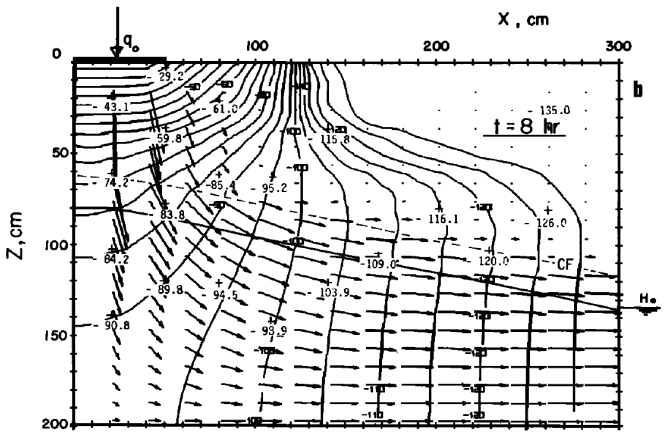


(a) Data from Vaulin et al. [43]

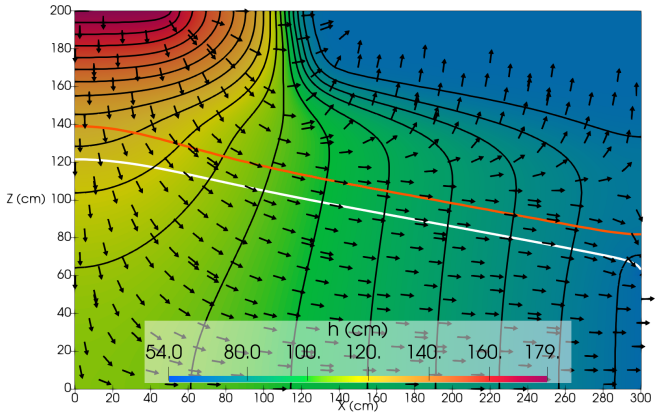


(b) Numerical solution

Figure 14: Vauclin's test case, at $t = 3\ h$, spacial distribution of hydraulic head, water table position (white line), contour plot of hydraulic head (red lines) and flux (arrows).

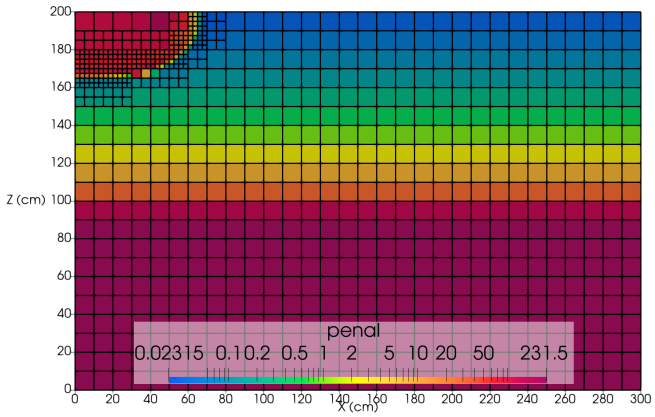


(a) Data from Vaulin et al. [43]

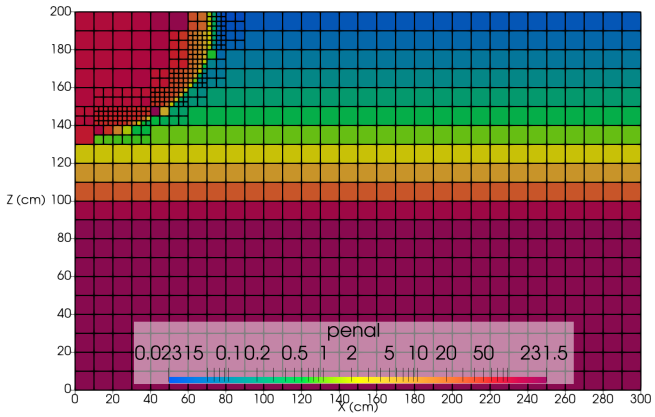


(b) Numerical solution

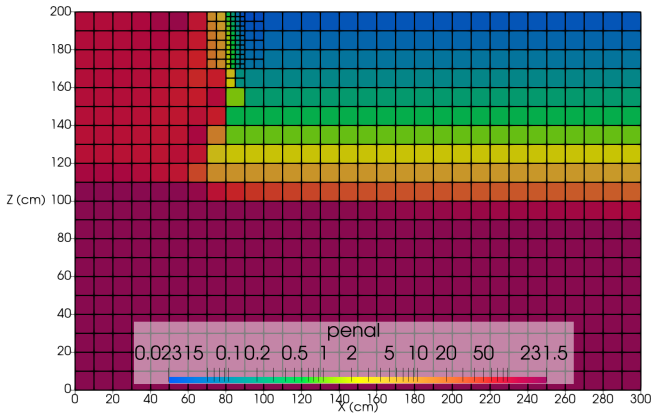
Figure 15: Vauclin's test case, at $t = 8\ h$ spacial distribution of hydraulic head, water table position (white line), contour plot of hydraulic head (red lines) and flux (arrows).



(a) At $t = 0.5\ h$



(b) At $t = 1\ h$



(c) At $t = 2\ h$

Figure 16: Vauclin's test case, spatial distribution of penalization parameters and mesh at selected times.

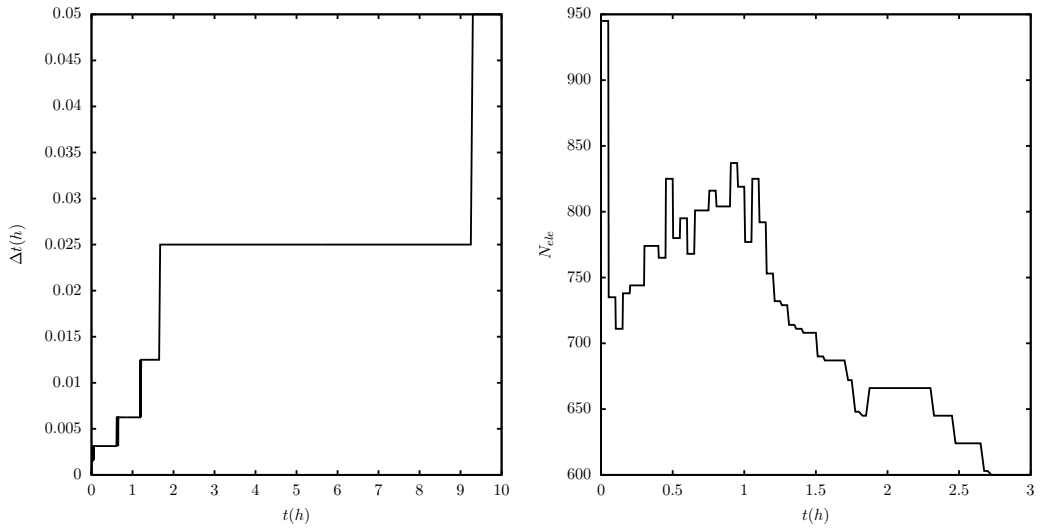


Figure 17: Vauclin's test case, evolution along time of time-steps (left) and number of elements (right).

An upper bound to the term $\sum_{n=0}^{N} \{\!\{K(x,\bar{u})u_h'\}\!\}_{x_n} [\![u_h]\!]_{x_n}$ needs to be established to prove the coercivity of $\tilde{a}_h$. Using Hypothesis $(\mathcal{H}_n)$ and definition of average:

$$\begin{aligned}
&\forall n \in \{1, ..., N-1\} \\
&|\{\!\{K(x,\bar{u})u_h'\}\!\}_{x_n}| \leq \frac{1}{2}\Big(|K(x_n^-, \bar{u}(x_n^-))u_h'(x_n^-)| + |K(x_n^+, \bar{u}(x_n^+))u_h'(x_n^+)|\Big) \\
&\leq \frac{K_1^{(n-1)}}{2}|u_h'(x_n^-)| + \frac{K_1^{(n)}}{2}|u_h'(x_n^+)|
\end{aligned}$$

Recalling the trace inequality [44] in the case of an orthonormal polynomial basis: for an interval I_n,

$$\forall u \in \mathbb{P}^p(I_n), \quad |u(x_n^+)| \leq C_{\mathrm{tr},p}\frac{\|u\|_{L^2(I_n)}}{\sqrt{h}}, \quad |u(x_{n+1}^-)| \leq C_{\mathrm{tr},p}\frac{\|u\|_{L^2(I_n)}}{\sqrt{h}}$$

we get, $\forall n \in \{1, ..., N-1\}$:

$$\begin{aligned}
|\{\!\{K(x,\bar{u})u_h'\}\!\}_{x_n}||[\![u_h]\!]_{x_n}| &\leq \Big(\frac{K_1^{(n-1)}}{2}\frac{C_{\mathrm{tr},p-1}^{(n-1)}}{\sqrt{h}}\|u_h'\|_{I_{n-1}} + \frac{K_1^{(n)}}{2}\frac{C_{\mathrm{tr},p-1}^{(n)}}{\sqrt{h}}\|u_h'\|_{I_n}\Big)|[\![u_h]\!]_{x_n}| \\
&\leq \sqrt{\varepsilon^{(n-1)}}\sqrt{K_0^{(n-1)}}\|u_h'\|_{I_{n-1}}\frac{K_1^{(n-1)}}{2\sqrt{\varepsilon^{(n-1)}}\sqrt{K_0^{(n-1)}}}\frac{C_{\mathrm{tr},p-1}^{(n-1)}}{\sqrt{h}}|[\![u_h]\!]_{x_n}| \\
&+ \sqrt{\varepsilon^{(n)}}\sqrt{K_0^{(n)}}\|u_h'\|_{I_n}\frac{K_1^{(n)}}{2\sqrt{\varepsilon^{(n)}}\sqrt{K_0^{(n)}}}\frac{C_{\mathrm{tr},p-1}^{(n)}}{\sqrt{h}}|[\![u_h]\!]_{x_n}|
\end{aligned}$$

At the boundary nodes x_0 and x_N, we have

$$|\{\!\{K(x,\bar{u})u_h'\}\!\}_{x_0}||[\![u_h]\!]_{x_0}| \leq \sqrt{\varepsilon^{(0)}}\sqrt{K_0^{(0)}}\|u_h'\|_{I_0}\frac{K_1^{(0)}}{\sqrt{\varepsilon^{(0)}}\sqrt{K_0^{(0)}}}\frac{C_{\mathrm{tr},p-1}^{(0)}}{\sqrt{h}}|[\![u_h]\!]_{x_0}|$$

$$|\{\!\{K(x,\bar{u})u_h'\}\!\}_{x_N}||[\![u_h]\!]_{x_N}| \leq \sqrt{\varepsilon^{(N-1)}}\sqrt{K_0^{(N-1)}}\|u_h'\|_{I_{N-1}}\frac{K_1^{(N-1)}}{\sqrt{\varepsilon^{(N-1)}}\sqrt{K_0^{(N-1)}}}\frac{C_{\mathrm{tr},p-1}^{(N-1)}}{\sqrt{h}}|[\![u_h]\!]_{x_N}|$$

Gathering the bounds on the boundary and the interior nodes, we get

$$\begin{aligned}\sum_{n=0}^{N-1} |\{\!\{K(x,\bar{u})u_h'\}\!\}_{x_n}||[\![u_h]\!]_{x_n}| \leq \sqrt{\varepsilon^{(0)}}\sqrt{K_0^{(0)}}\|u_h'\|_{I_0}\frac{K_1^{(0)}}{\sqrt{\varepsilon^{(0)}}\sqrt{K_0^{(0)}}}\frac{C_{\mathrm{tr},p-1}^{(0)}}{\sqrt{h}}|[\![u_h]\!]_{x_0}| \\ + \sum_{n=1}^{N-1}\Bigg(\sqrt{\varepsilon^{(n-1)}}\sqrt{K_0^{(n-1)}}\|u_h'\|_{I_{n-1}}\frac{K_1^{(n-1)}}{2\sqrt{\varepsilon^{(n-1)}}\sqrt{K_0^{(n-1)}}}\frac{C_{\mathrm{tr},p-1}^{(n-1)}}{\sqrt{h}}|[\![u_h]\!]_{x_n}| \\ + \sqrt{\varepsilon^{(n)}}\sqrt{K_0^{(n)}}\|u_h'\|_{I_n}\frac{K_1^{(n)}}{2\sqrt{\varepsilon^{(n)}}\sqrt{K_0^{(n)}}}\frac{C_{\mathrm{tr},p-1}^{(n)}}{\sqrt{h}}|[\![u_h]\!]_{x_n}|\Bigg) \\ + \sqrt{\varepsilon^{(N-1)}}\sqrt{K_0^{(N-1)}}\|u_h'\|_{I_{N-1}}\frac{K_1^{(N-1)}}{\sqrt{\varepsilon^{(N-A)}}\sqrt{K_0^{(N-1)}}}\frac{C_{\mathrm{tr},p-1}^{(N-1)}}{\sqrt{h}}|[\![u_h]\!]_{x_N}|\end{aligned}$$

Then, using Cauchy-Schwarz's and Young's inequality, we have:

$$\begin{aligned}&\sum_{n=0}^{N-1}\{\!\{K(x,\bar{u})u_h'\}\!\}_{x_n}[\![u_h]\!]_{x_n} \leq \\ &\sum_{n=0}^{N-1}\frac{\varepsilon^{(n)}K_0^{(n)}}{2}\|u_h'\|_{I_n}^2 + \frac{(K_1^{(0)}C_{\mathrm{tr},p-1}^{(0)})^2}{\varepsilon^{(0)}K_0^{(0)}}\frac{[\![u_h]\!]_{x_0}^2}{h} + \frac{(K_1^{(N-1)}C_{\mathrm{tr},p-1}^{(N-1)})^2}{\varepsilon^{(N-1)}K_0^{(N-1)}}\frac{[\![u_h]\!]_{x_N}^2}{h} \\ &+ \sum_{n=1}^{N-1}\Big(\frac{(K_1^{(n-1)}C_{\mathrm{tr},p-1}^{(n-1)})^2}{2\varepsilon^{(n-1)}K_0^{(n-1)}}\frac{[\![u_h]\!]_{x_n}^2}{2h} + \frac{(K_1^{(n)}C_{\mathrm{tr},p-1}^{(n)})^2}{2\varepsilon^{(n)}K_0^{(n)}}\frac{[\![u_h]\!]_{x_n}^2}{2h}\Big)\end{aligned}$$

From the above inequality, we deduce a lower bound of $\tilde{a}_h(u_h,u_h;\bar{u}),\ \forall u_h \in \mathcal{V}_0^p(\mathcal{E}_h)$

$$\begin{aligned}\tilde{a}_h(u_h,u_h) \geq \sum_{n=0}^{N-1}\Big(K_0^{(n)} - \frac{\varepsilon^{(n)}K_0^{(n)}}{2}\Big)\|u_h'\|_{I_n}^2 + \sum_{n=1}^{N-1}\frac{(\sigma_{n-1}-\sigma_{n-1}^*)+(\sigma_n-\sigma_n^*)}{2}\frac{1}{h}[\![u_h]\!]_{x_n}^2 \\ + (\sigma_0-\sigma_0^*)\frac{1}{h}[\![u_h]\!]_{x_0}^2 + (\sigma_N-\sigma_N^*)\frac{1}{h}[\![u_h]\!]_{x_N}^2 \qquad (15)\end{aligned}$$

where

$$\begin{cases}\sigma_n^* = \dfrac{(K_1^{(n)}C_{\mathrm{tr},p-1}^{(n)})^2}{2\varepsilon^{(n)}K_0^{(n)}},\ \forall n \in \{1,...,N-1\} \\ \sigma_0^* = \dfrac{(K_1^{(0)}C_{\mathrm{tr},p-1}^{(0)})^2}{\varepsilon^{(0)}K_0^{(0)}} \\ \sigma_N^* = \dfrac{(K_1^{(N-1)}C_{\mathrm{tr},p-1}^{(N-1)})^2}{\varepsilon^{(N-1)}K_0^{(N-1)}}\end{cases}$$

Finally, thanks to the inequality (15), $\tilde{a}_h$ (11) is coercive if

$$\begin{cases}\varepsilon^{(n)} < 2, & \forall n \in \{0,...,N-1\} \\ \sigma_n > \sigma_n^*, & \forall n \in \{1,...,N-1\} \\ \sigma_0 > \sigma_0^* & \\ \sigma_N > \sigma_N^* & \end{cases}$$

which ends the proof. □

Proof of Lemma 3.4. For a given $\bar{u} \in \mathcal{V}_0^p(\mathcal{E}_h)$, an upper bound for $|\tilde{a}_h(u_h, v_h; \bar{u})|$, $\forall u_h, v_h \in \mathcal{V}_0^p(\mathcal{E}_h)$ needs to be established in order to prove continuity of $\tilde{a}_h$. Firstly, start bounding above the volume contribution using Hypothesis $(\mathcal{H}_n)$:

$$\begin{aligned}\left|\sum_{n=0}^{N-1}\int_{I_n} K(x,\bar{u})u_h'v_h'dx\right| &\leq \sum_{n=0}^{N-1} K_1^{(n)}\left|\int_{I_n} u_h'v_h'dx\right| \leq \sum_{n=0}^{N-1}\sqrt{K_1^{(n)}}\|u_h'\|_{I_n}\sqrt{K_1^{(n)}}\|v_h'\|_{I_n}\\ &\leq \Big(\sum_{n=0}^{N-1} K_1^{(n)}\|u_h'\|_{I_n}^2\Big)^{\frac{1}{2}}\Big(\sum_{n=0}^{N-1} K_1^{(n)}\|v_h'\|_{I_n}^2\Big)^{\frac{1}{2}}.\end{aligned}$$

Then, penalization terms are bounded above

$$\begin{aligned}&\left|\frac{\sigma_0}{h}[\![u_h]\!]_{x_0}[\![v_h]\!]_{x_0} + \sum_{n=1}^{N-1}\frac{\sigma_{n-1}+\sigma_n}{2h}[\![u_h]\!]_{x_n}[\![v_h]\!]_{x_n} + \frac{\sigma_N}{h}[\![u_h]\!]_{x_N}[\![v_h]\!]_{x_N}\right|\\ \leq&\Big(\frac{\sigma_0}{h}[\![u_h]\!]_{x_0}^2 + \sum_{n=1}^{N-1}\frac{\sigma_{n-1}+\sigma_n}{2h}[\![u_h]\!]_{x_n}^2 + \frac{\sigma_N}{h}[\![u_h]\!]_{x_N}^2\Big)^{\frac{1}{2}}\\ &\Big(\frac{\sigma_0}{h}[\![v_h]\!]_{x_0}^2 + \sum_{n=1}^{N-1}\frac{\sigma_{n-1}+\sigma_n}{2h}[\![v_h]\!]_{x_n}^2 + \frac{\sigma_N}{h}[\![v_h]\!]_{x_N}^2\Big)^{\frac{1}{2}}\\ \leq&\max\left(\sigma_0,\sigma_N,\max_{n=1,\dots,N-1}\Big(\frac{\sigma_n}{2}\Big)\right)\|u_h\|\|v_h\|\end{aligned}$$

and one can write

$$\begin{aligned}\sum_{n=0}^{N}|\{\!\{K(x,\bar{u})u_h'\}\!\}_{x_n}[\![v_h]\!]_{x_n}| \leq&\left(2\sum_{n=0}^{N-1}\varepsilon^{(n)}K_0^{(n)}\|u_h'\|_{I_n}^2\right)^{\frac{1}{2}}\\ &\left(\frac{(K_1^{(0)}C_{\mathrm{tr},p-1}^{(0)})^2}{\varepsilon^{(0)}K_0^{(0)}}\frac{[\![v_h]\!]_{x_0}^2}{h} + \frac{(K_1^{(N-1)}C_{\mathrm{tr},p-1}^{(N-1)})^2}{\varepsilon^{(N-1)}K_0^{(N-1)}}\frac{[\![v_h]\!]_{x_N}^2}{h}\right.\\ &\left.+\sum_{n=1}^{N-1}\Big(\frac{(K_1^{(n-1)}C_{\mathrm{tr},p-1}^{(n-1)})^2}{2\varepsilon^{(n-1)}K_0^{(n-1)}} + \frac{(K_1^{(n)}C_{\mathrm{tr},p-1}^{(n)})^2}{2\varepsilon^{(n)}K_0^{(n)}}\Big)\frac{[\![v_h]\!]_{x_n}^2}{2h} + \right)^{\frac{1}{2}}.\end{aligned}$$

From those inequalities, we obtain an upper bound $\forall u_h, v_h \in \mathcal{V}_0^p(\mathcal{E}_h)$, as follows

$$\begin{aligned}|\tilde{a}_h(u_h,v_h;\bar{u})| \leq&\Big(\sum_{n=0}^{N-1}K_1^{(n)}\|u_h'\|_{I_n}^2\Big)^{\frac{1}{2}}\Big(\sum_{n=0}^{N-1}K_1^{(n)}\|v_h'\|_{I_n}^2\Big)^{\frac{1}{2}}\\ &+\left(\sum_{n=0}^{N-1}2\varepsilon^{(n)}K_1^{(n)}\|u_h'\|_{I_n}^2\right)^{\frac{1}{2}}\left(\frac{\sigma_0^*}{h}[\![v_h]\!]_{x_0}^2 + \sum_{n=1}^{N-1}\frac{\sigma_{n-1}^*+\sigma_n^*}{2h}[\![v_h]\!]_{x_n}^2 + \frac{\sigma_N^*}{h}[\![v_h]\!]_{x_N}^2\right)^{\frac{1}{2}}\\ &+\max\left(\sigma_0,\sigma_N,\max_{n=1,\dots,N-1}\Big(\frac{\sigma_n}{2}\Big)\right)\|u_h\|\|v_h\|\\ \leq&\max_{n=0,\dots,N-1}\left(K_1^{(n)}\right)\|u_h\|\|v_h\|\\ &+\sqrt{\max_{n=0,\dots,N-1}\left(2\varepsilon^{(n)}K_1^{(n)}\right)\max\left(\sigma_0^*,\sigma_N^*,\max_{n=1,\dots,N-1}\left(\frac{\sigma_n^*}{2}\right)\right)}\|u_h\|\|v_h\|\\ &+\max\left(\sigma_0,\sigma_N,\max_{n=1,\dots,N-1}\Big(\frac{\sigma_n}{2}\Big)\right)\|u_h\|\|v_h\|\\ \leq&\tilde{C}\|u_h\|\|v_h\|\end{aligned}$$

where

$$\tilde{C}(\epsilon) = \max_{n=0,\dots,N-1}\left(K_1^{(n)}\right) + \sqrt{\max_{n=0,\dots,N-1}\left(2\varepsilon^{(n)}K_1^{(n)}\right)\max\left(\sigma_0^*, \sigma_N^*, \max_{n=1,\dots,N-1}\left(\frac{\sigma_n^*}{2}\right)\right)}$$
$$+ \max\left(\sigma_0, \sigma_N, \max_{n=1,\dots,N-1}\left(\frac{\sigma_n}{2}\right)\right).$$

□

Proof of Lemma 3.5. An upper bound for $|l(v_h)|$ is established using Poincaré inequality and Cauchy Schwarz: $\forall v_h \in \mathcal{V}_0^p(\mathcal{E}_h)$,

$$\left|\sum_{n=0}^{N-1}\int_{I_n} f v dx\right| \leq \sum_{n=0}^{N-1}\|f\|_{I_n}\|v_h\|_{I_n} \leq \sum_{n=0}^{N-1}\|f\|_{I_n}\beta_n\|v_h'\|_{I_n}$$
$$\leq \Big(\sum_{n=0}^{N-1}\|f\|_{I_n}^2\Big)^{\frac{1}{2}}\Big(\sum_{n=0}^{N-1}\beta_n^2\|v_h'\|_{I_n}^2\Big)^{\frac{1}{2}} \leq B\|v_h\|$$

with $B = \max_{n=0,\dots,N-1}(\beta_n)\Big(\sum_{n=0}^{N-1}\|f\|_{I_n}^2\Big)^{\frac{1}{2}}$. □

Proof of Lemma 3.9. For a given $\bar{u} \in \mathcal{V}_0^p(\mathcal{E}_h)$, let v_h be a sequence in $\mathcal{V}_0^p(\mathcal{E}_h)$ bounded in the $\|.\|$-norm and let $\varphi \in C_0^\infty(\Omega)$. For all $h \in \mathbb{R}_+^*$, set $\varphi_h = \pi_h\varphi$ where π_h denotes the L^2-orthogonal projection onto $\mathcal{V}_0^p(\mathcal{E}_h)$. Since $p \geq 1$, infer $\|\varphi - \pi_h\varphi\| \xrightarrow[h\to 0]{} 0$. Owing to Equation (14) and since $G_h^p(\varphi) = \varphi'$ because $\varphi \in C_0^\infty(\Omega)$, obtain for all $p \geq 0$

$$G_h^p(\pi_h\varphi) \longrightarrow \varphi' \text{ strongly in } L^2(\Omega)$$

One can observe that

$$\tilde{a}_h(v_h, \pi_h\varphi; \bar{u}) = \int_\Omega K(x,\bar{u})\hat{G}_h^p(v_h)G_h^p(\pi_h\varphi)dx + s_h(v_h, \pi_h\varphi) := \mathfrak{T}_1 + \mathfrak{T}_2$$

Clearly as $h \to 0$,$\mathfrak{T}_1 \to \int_\Omega K(x,\bar{u})v'\varphi' dx$ owing to the weak convergence of $\hat{G}_h^p(v_h)$ to v' and to the strong convergence of $G_h^p(\pi_h\varphi)$ to φ'. Furthermore, using Cauchy-Schwarz inequality yields :

$$|\mathfrak{T}_2| = |s_h(v_h, \pi_h\varphi)|$$
$$= |\frac{\sigma_0}{h}[\![v_h]\!]_{x_0}[\![\pi_h\varphi]\!]_{x_0} + \sum_{n=1}^{N-1}\frac{\sigma_{n-1}+\sigma_n}{2h}[\![v_h]\!]_{x_n}[\![\pi_h\varphi]\!]_{x_n} + \frac{\sigma_N}{h}[\![v_h]\!]_{x_N}[\![\pi_h\varphi]\!]_{x_N}|$$
$$\leq \Big(\sigma_0^2\frac{[\![v_h]\!]_{x_0}^2}{h} + \sum_{n=1}^{N-1}\frac{(\sigma_{n-1}+\sigma_n)^2}{4}\frac{[\![v_h]\!]_{x_n}^2}{h} + \sigma_N^2\frac{[\![v_h]\!]_{x_N}^2}{h}\Big)^{\frac{1}{2}}\Big(\sum_{n=0}^{N}\frac{[\![\pi_h\varphi]\!]_{x_n}^2}{h}\Big)^{\frac{1}{2}}$$
$$\leq \mathcal{C}|v_h|_J|\pi_h\varphi|_J$$

where

$$\mathcal{C} = \max\Big\{\sigma_0^2, \frac{(\sigma_{n-1}+\sigma_n)^2}{4}, \sigma_N^2\Big\}.$$

Since $|v_h|_J$ is bounded by assumption and since $|\pi_h\varphi|_J = |\varphi - \pi_h\varphi|_J \xrightarrow[h\to 0]{} 0$, infer $\mathfrak{T}_2 \xrightarrow[h\to 0]{} 0$. □

Proof of the Theorem 3.4. For a given $\bar{u} \in \mathcal{V}_0^p(\mathcal{E}_h)$, owing to the discrete coercivity of $\tilde{a}_h$, the sequence u_h is bounded in the $\|.\|$-norm. Theorem 3.3 implies that there is $v \in H_0^1(\Omega)$ such that up to a subsequence,

$u_h \to v$ in $L^2(\Omega)$ and for all $p \geq 0$, $G_h^p(u_h) \rightharpoonup v'$ weakly in $L^2(\Omega)$ as $h \to 0$. Let $\varphi \in C_0^\infty(\Omega)$. Owing Lemma 3.9, $\tilde{a}_h(u_h, \pi_h\varphi; \bar{u}) \to \tilde{a}(v, \varphi)$ as $h \to 0$. Since u_h solves the discrete linearized problem $(\tilde{\mathcal{W}}_h)$, infer as $h \to 0$

$$\begin{aligned}\tilde{a}_h(u_h - \pi_h\varphi, u_h - \pi_h\varphi; \bar{u}) &= \tilde{a}_h(u_h, u_h - \pi_h\varphi; \bar{u}) - \tilde{a}_h(\pi_h\varphi, u_h - \pi_h\varphi; \bar{u}) \\ &\longrightarrow \tilde{a}(v, v - \varphi) - \tilde{a}(\varphi, v - \varphi) \\ &\longrightarrow \int_\Omega (v - \varphi) f dx - \int_\Omega K(x, \bar{u}) \varphi'(v - \varphi)' dx\end{aligned}$$

Hence using $\tilde{a}_h(v_h, v_h; \bar{u}) \geq C^* \|v_h\|^2$ from Lemma 3.3

$$\begin{aligned}C^* \|u_h - \pi_h\varphi\| &\leq \tilde{a}_h(u_h - \pi_h\varphi, u_h - \pi_h\varphi; \bar{u}) \\ \Leftrightarrow \limsup_{h\to 0} C^* \|u_h - \pi_h\varphi\| &\leq \limsup_{h\to 0} \tilde{a}_h(u_h - \pi_h\varphi, u_h - \pi_h\varphi; \bar{u}) \\ &\leq |\int_\Omega (v - \varphi) f dx - \int_\Omega K(x, \bar{u}) \varphi'(v - \varphi)' dx| \\ &\leq \|f\|_{L^2(\Omega)} \|v - \varphi\|_{L^2(\Omega)} + K_1 \|\varphi'\|_{L^2(\Omega)} \|(v - \varphi)'\|_{L^2(\Omega)} \\ &\leq C_{f,\varphi} \Big(\|v - \varphi\|^2_{L^2(\Omega)} + \|(v - \varphi)'\|^2_{L^2(\Omega)} \Big)^{\frac{1}{2}} \\ &\leq C_{f,\varphi} \|v - \varphi\|_{H^1(\Omega)}\end{aligned}$$

with $C_{f,\varphi} = \Big(\|f\|^2_{L^2(\Omega)} + K_1 \|\varphi'\|^2_{L^2(\Omega)} \Big)^{\frac{1}{2}}$. As a consequence

$$\limsup_{h\to 0} \|u_h - \pi_h\varphi\| \leq \frac{1}{C^*} C_{f,\varphi} \|v - \varphi\|_{H^1(\Omega)}.$$

One can observe that the choice for $\hat{G}_h^p$ satisfy the stability property

$$\forall v_h \in \mathcal{V}_0^p(\mathcal{E}_h), \quad \|\hat{G}_h^p(v_h)\|_{L^2(\Omega)} \leq \hat{C} \|v_h\|$$

for $\hat{C}$ independent of h. As a result,

$$\limsup_{h\to 0} \|\hat{G}_h^p(u_h) - \hat{G}_h^p(\pi_h\varphi)\|_{L^2(\Omega)} \leq \hat{C} \frac{1}{C^*} C_{f,\varphi} \|v - \varphi\|_{H^1(\Omega)}$$

because

$$\begin{aligned}\|\hat{G}_h^p(u_h) - \hat{G}_h^p(\pi_h\varphi)\|_{L^2(\Omega)} &\leq \hat{C} \|u_h - \pi_h\varphi\| \\ \Leftrightarrow \limsup_{h\to 0} \|\hat{G}_h^p(u_h) - \hat{G}_h^p(\pi_h\varphi)\|_{L^2(\Omega)} &\leq \hat{C} \limsup_{h\to 0} \|u_h - \pi_h\varphi\| \\ &\leq \hat{C} \frac{1}{C^*} C_{f,\varphi} \|v - \varphi\|_{H^1(\Omega)}\end{aligned}$$

And since $\hat{G}_h^p(\pi_h\varphi)$ strongly converges to φ' in $L^2(\Omega)$, this yields

$$\limsup_{h\to 0} \|\hat{G}_h^p(u_h) - \varphi'\|_{L^2(\Omega)} \leq \hat{C} \frac{1}{C^*} C_{f,\varphi} \|v - \varphi\|_{H^1(\Omega)}.$$

Since φ is arbitrary in $C_0^\infty(\Omega)$, and since this space is dense in $H_0^1(\Omega)$, the term on the right hand side can be made as small as desired taking $\varphi = v$, infer

$$\hat{G}_h^p(u_h) \xrightarrow[h\to 0]{} v' \text{ strongly in } L^2(\Omega)$$

As a result, taking φ arbitrary in $C_0^\infty(\Omega)$ yields

$$\int_\Omega K(x,\bar{u})v'\varphi' dx \xleftarrow[h\to 0]{} \int_\Omega K(x,\bar{u})u_h'\pi_h\varphi' dx = \tilde{a}_h(u_h,\pi_h\varphi) = \int_\Omega f\pi_h\varphi \xrightarrow[h\to 0]{} \int_\Omega f\varphi dx \tag{16}$$

using Lemma 3.9. i.e. v solves the poisson problem by density of $C_0^\infty(\Omega)$ in $H^1(\Omega)$. Since the solution u to the Poisson problem is unique, the whole sequence u_h strongly converges to u in $L^2(\Omega)$ and, for all $p \geq 0$, the sequence $\left(G_h^p(u_h)\right)_{h\in\mathbb{R}_+^*}$ weakly converges to u' in $L^2(\Omega)$.

$$\begin{aligned}\tilde{a}_h(u_h,u_h;\bar{u}) &= \int_\Omega K(x,\bar{u})\hat{G}_h^p(u_h)G_h^p(u_h)dx + s_h(u_h,u_h)\\ &\geq \int_\Omega K(x,\bar{u})\hat{G}_h^p(u_h)G_h^p(u_h)dx\end{aligned}$$

Thus

$$\liminf_{h\to 0}\tilde{a}_h(u_h,u_h;\bar{u}) \geq \liminf_{h\to 0}\int_\Omega K(x,\bar{u})\hat{G}_h^p(u_h)G_h^p(u_h)dx \geq \int_\Omega K(x,\bar{u})u'u'dx$$

Furthermore

$$\int_\Omega K(x,\bar{u})\hat{G}_h^p(u_h)G_h^p(u_h)dx \leq \tilde{a}_h(u_h,u_h;\bar{u}) = \int_\Omega f u_h dx$$

yielding with Equation (16)

$$\begin{aligned}\limsup_{h\to 0}\int_\Omega K(x,\bar{u})\hat{G}_h^p(u_h)G_h^p(u_h)dx &\leq \limsup_{h\to 0}\tilde{a}_h(u_h,u_h;\bar{u})\\ &= \limsup_{h\to 0}\int_\Omega f u_h dx \leq \int_\Omega K(x,\bar{u})u'u'dx\end{aligned}$$

Thus, $\int_\Omega K(x,\bar{u})\hat{G}_h^p(u_h)G_h^p(u_h)dx \xrightarrow[h\to 0]{} \int_\Omega K(x,\bar{u})u'u'dx$ strongly. Moreover $\tilde{a}_h(u_h,u_h;\bar{u}) \xrightarrow[h\to 0]{} \int_\Omega K(x,\bar{u})u'u'dx$ strongly. Owing that

$$\begin{aligned}\tilde{a}_h(u_h,u_h;\bar{u}) &= \int_\Omega K(x,\bar{u})\hat{G}_h^p(u_h)G_h^p(u_h)dx + s_h(u_h,u_h)\\ &\geq \int_\Omega K(x,\bar{u})\hat{G}_h^p(u_h)G_h^p(u_h)dx + \min_{n=0,\dots,N}(\sigma_n)|u_h|_J^2\\ \Leftrightarrow \min_{n=0,\dots,N}(\sigma_n)|u_h|_J^2 &\leq a_h(u_h,u_h) - \int_\Omega K(x,\bar{u})\hat{G}_h^p(u_h)G_h^p(u_h)dx\end{aligned}$$

and since $\min_{n=0,\dots,N}(\sigma_n) > 0$ and the right-hand side tends to zero, $|u_h|_J \to 0$.

$$\|u_h' - u'\|_{L^2(\Omega)} = \|\hat{G}_h^p(u_h) - u'\|_{L^2(\Omega)} \longrightarrow 0,$$

□

References

[1] T. Altazin, M. Ersoy, F. Golay, D. Sous, and L. Yushchenko. Numerical investigation of BB-AMR scheme using entropy production as refinement criterion. *International Journal of Computational Fluid Dynamics*, 30(3):256–271, Mar. 2016.

[2] F. Bassi and S. Rebay. A High-Order Accurate Discontinuous Finite Element Method for the Numerical Solution of the Compressible Navier–Stokes Equations. *Journal of Computational Physics*, 131(2):267–279, 1997.

[3] L. Bergamaschi and M. Putti. Mixed finite elements and Newton-type linearizations for the solution of Richards' equation. *International Journal for Numerical Methods in Engineering*, 45(8):1025–1046, 1999.

[4] L. Boccardo, G. Thierry, and F. Murat. Unicité de la solution de certaines équations elliptiques non linéaires. *C. R. Acad. Sci. Paris*, 315:1159–1164, Jan. 1992.

[5] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous Galerkin approximations for elliptic problems. *Numerical Methods for Partial Differential Equations*, 16(4):365–378, 2000.

[6] E. Buckingham. *Studies on the Movement of Soil Moisture*, volume 38. Washington, Govt. Print. Off., bulletin edition, 1907.

[7] M. A. Celia, E. T. Bouloutas, and R. L. Zarba. A general mass-conservative numerical solution for the unsaturated flow equation. *Water Resources Research*, 26(7):1483–1496, July 1990.

[8] J.-B. Clément. *Numerical Simulation of Flows in Unsaturated Porous Media by an Adaptive Discontinuous Galerkin Method: Application to Sandy Beaches*. PhD thesis, Université de Toulon, Jan. 2021.

[9] J.-B. Clément, F. Golay, M. Ersoy, and D. Sous. An adaptive strategy for discontinuous Galerkin simulations of Richards' equation: Application to multi-materials dam wetting. *Advances in Water Resources*, 151:103897, 2021.

[10] R. L. Cooley. A Finite Difference Method for Unsteady Flow in Variably Saturated Porous Media: Application to a Single Pumping Well. *Water Resources Research*, 7(6):1607–1625, Dec. 1971.

[11] G. G. Dahlquist. A special stability problem for linear multistep methods. *BIT*, 3(1):27–43, Mar. 1963.

[12] H. Darcy. *Les Fontaines Publiques de La Ville de Dijon*. Librairie des corps impériaux des ponts et des chaussées et des mines, 1856.

[13] A. Dogan and L. H. Motz. Saturated-Unsaturated 3D Groundwater Model. II: Verification and Application. *Journal of Hydrologic Engineering*, 10(6):505–515, Nov. 2005.

[14] V. Dolejší and M. Feistauer. *Discontinuous Galerkin Method*. Springer International Publishing, 2015.

[15] V. Dolejší, M. Kuraz, and P. Solin. Adaptive higher-order space-time discontinuous Galerkin method for the computer simulation of variably-saturated porous media flows. *Applied Mathematical Modelling*, 72:276–305, Aug. 2019.

[16] Y. Epshteyn and B. Rivière. Estimation of penalty parameters for symmetric interior penalty Galerkin methods. *Journal of Computational and Applied Mathematics*, 206(2):843–872, 2007.

[17] M. Ersoy, F. Golay, and L. Yushchenko. Adaptive multiscale scheme based on numerical density of entropy production for conservation laws. *Open Mathematics*, 11(8), Jan. 2013.

[18] M. W. Farthing, C. E. Kees, and C. T. Miller. Mixed finite element methods and higher order temporal approximations for variably saturated groundwater flow. *Advances in Water Resources*, 26(4):373–394, Apr. 2003.

[19] M. W. Farthing and F. L. Ogden. Numerical Solution of Richards' Equation: A Review of Advances and Challenges. *Soil Science Society of America Journal*, 81(6):1257–1269, Nov. 2017.

[20] F. Golay, M. Ersoy, L. Yushchenko, and D. Sous. Block-based adaptive mesh refinement scheme using numerical density of entropy production for three-dimensional two-fluid flows. *International Journal of Computational Fluid Dynamics*, 29(1):67–81, Jan. 2015.

[21] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II*, volume 14 of *Springer Series in Computational Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1996.

[22] R. Haverkamp, M. Vauclin, J. Touma, P. J. Wierenga, and G. Vachaud. A Comparison of Numerical Simulation Models For One-Dimensional Infiltration. *Soil Science Society of America Journal*, 41(2):285, 1977.

[23] A. Hay, S. Etienne, D. Pelletier, and A. Garon. Hp-Adaptive time integration based on the BDF for viscous flows. *Journal of Computational Physics*, 291:151–176, June 2015.

[24] S. Irmay. On the hydraulic conductivity of unsaturated soils. *Eos, Transactions American Geophysical Union*, 35(3):463–467, 1954.

[25] F. Lehmann and P. Ackerer. Comparison of Iterative Methods for Improved Solutions of the Fluid Flow Equation in Partially Saturated Porous Media. *Transport in Porous Media*, 31(3):275–292, 1998.

[26] H. Li, M. Farthing, C. Dawson, and C. Miller. Local discontinuous Galerkin approximations to Richards' equation. *Advances in Water Resources*, 30(3):555–575, Mar. 2007.

[27] F. List and F. A. Radu. A study on iterative methods for solving Richards' equation. *Computational Geosciences*, 20(2):341–353, Apr. 2016.

[28] G. Manzini and S. Ferraris. Mass-conservative finite volume methods on 2-D unstructured grids for the Richards' equation. *Advances in Water Resources*, 27(12):1199–1215, Dec. 2004.

[29] C. T. Miller, C. Abhishek, and M. W. Farthing. A spatially and temporally adaptive solution of Richards' equation. *Advances in Water Resources*, 29(4):525–545, Apr. 2006.

[30] J. R. Philip. THE THEORY OF INFILTRATION: 1. THE INFILTRATION EQUATION AND ITS SOLUTION. *Soil Science*, 171(6):S34, June 2006.

[31] D. A. D. Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*. Springer Berlin Heidelberg, 2012.

[32] L. A. Richards. Capillary conduction of liquids through porous mediums. *Physics*, 1(5):318–333, Nov. 1931.

[33] L. F. Richardson. *Weather Prediction by Numerical Process*. Cambridge University Press, 1922.

[34] B. Rivière. *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations*. Society for Industrial and Applied Mathematics, Jan. 2008.

[35] J. Rubin. Theoretical Analysis of Two-Dimensional, Transient Flow of Water in Unsaturated and Partly Unsaturated Soils. *Soil Science Society of America Journal*, 32(5):607–615, Sept. 1968.

[36] S. Skelboe. The control of order and steplength for backward differentiation methods. *BIT*, 17(1):91–107, Mar. 1977.

[37] P. Sochala. *Méthodes Numériques Pour Les Écoulements Souterrains et Couplage Avec Le Ruissellement*. PhD thesis, Dec. 2008.

[38] E. Süli and D. F. Mayers. *An Introduction to Numerical Analysis*. Cambridge University Press, 1 edition, Aug. 2003.

[39] R. B. Thoms, R. L. Johnson, and R. W. Healy. User's guide to the Variably Saturated Flow (VSF) process to MODFLOW, 2006.

[40] N. K. C. Twarakavi, J. Šimůnek, and S. Seo. Evaluating Interactions between Groundwater and Vadose Zone Using the HYDRUS-Based Flow Package for MODFLOW. *Vadose Zone Journal*, 7(2):757–768, 2008.

[41] G. Vachaud and J.-L. Thony. Hysteresis During Infiltration and Redistribution in a Soil Column at Different Initial Water Contents. *Water Resources Research*, 7(1):111–127, Feb. 1971.

[42] M. Th. van Genuchten. A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils. *Soil Science Society of America Journal*, 44(5):892–898, 1980.

[43] M. Vauclin, D. Khanji, and G. Vachaud. Experimental and numerical study of a transient, two-dimensional unsaturated-saturated water table recharge problem. *Water Resources Research*, 15(5):1089–1101, Oct. 1979.

[44] T. Warburton and J. S. Hesthaven. On the constants in hp-finite element trace inverse inequalities. *Computer Methods in Applied Mechanics and Engineering*, 192(25):2765–2773, 2003.

[45] Y. Zha, L. Shi, M. Ye, and J. Yang. A generalized Ross method for two- and three-dimensional variably saturated flow. *Advances in Water Resources*, 54:67–77, Apr. 2013.