Numerical methods around shallow water flows : dispersive effects, Coriolis force

Virgile Dubos

Sorbonne Université
Laboratoire Jacques-Louis Lions

Inria
Équipe-projet ANGE

École doctorale de sciences mathématiques de Paris centre

Thèse de Doctorat

Discipline : Sciences Mathématiques

présentée par

Virgile Dubos

Méthodes numériques autour d'écoulements en eaux peu profondes : effets dispersifs, force de Coriolis

---

sous la direction de Cindy Guichard, Yohan Penel et Jacques Sainte-Marie

soutenue le 14 décembre 2021 devant un jury composé de :

Mme	Claire Chainais-Hillairet	Professeure	Rapportrice
Mme	Cindy Guichard	Maîtresse de conférences	Encadrante
Mme	Raphaèle Herbin	Professeure	Examinatrice
M.	Pierre-Yves Lagrée	Directeur de recherche	Examinateur
M.	Pascal Omnes	Directeur de recherche	Président du jury
M.	Yohan Penel	Chargé de recherche	Encadrant
M.	Jacques Sainte-Marie	Directeur de recherche	Directeur de thèse

après avis favorable des rapporteurs :

Mme	Claire Chainais-Hillairet	Professeure
M.	Rainer Helmig	Professeur

Résumé

Nous étudions dans cette thèse des méthodes numériques pour les écoulements en eaux peu profondes à surface libre. D'une part, nous nous intéressons aux modèles non-hydrostatique de Saint-Venant prenant en compte des effets dispersifs. D'autre part, nous étudions le modèle de Saint-Venant avec terme source de Coriolis et ses équilibres géostrophiques.

Dans un premier temps nous considérons des méthodes numériques pour une famille de modèles d'Euler moyennés sur la hauteur d'eau provenant de la littérature. Il s'agit de modèles de Saint-Venant avec une pression non-hydrostatique. Les méthodes numériques considérées sont basées sur une méthode de prédiction-correction consistant à décomposer le problème en deux étapes à chaque itération en temps. L'étape de prédiction mène à résoudre un système de Saint-Venant qui est habituellement résolu par une méthode de volumes finis, alors que l'étape de correction mène à résoudre un problème elliptique régissant la pression non-hydrostatique. Notre but premier est d'analyser la convergence d'un schéma mixte avec condensation de masse introduit dans la littérature en utilisant la Méthode de Discrétisation du Gradient (GDM) qui forme un cadre englobant des schémas de discrétisation usuels et récents pour des problèmes de diffusion. Cette méthode nous permet d'obtenir la convergence du schéma proposé pour le problème elliptique. Par la suite, nous proposons une nouvelle formulation conforme du problème en pression et donnons les estimateurs d'erreur correspondant via la GDM. Enfin, nous donnons un exemple d'application avec la méthode des éléments finis $\mathbb{P}_1$.

Dans un second temps nous souhaitons avoir des schémas explicites colocalisés volumes finis pour les équations de Saint-Venant non-linéaires avec un terme source de Coriolis qui seraient précis aux alentours de l'équilibre géostrophique et stables dans le cadre non-linéaire. Nous construisons plusieurs schémas volumes finis et étudions les deux propriétés suivantes : la décroissance de l'énergie semi-discrète et la préservation de l'équilibre géostrophique par la version linéarisée. Nous proposons également une version conservative de l'un de nos schéma. Enfin, nous observons le comportement de ces schémas à travers plusieurs cas tests et obtenons de meilleurs résultats en comparaison d'un schéma volumes finis classique.

Mots-clés : écoulements à surface libre, équations de Saint-Venant, pression non-hydrostatique, méthode de discrétisation du gradient, éléments finis, force de Coriolis, équilibre géostrophique, schéma équilibre, diffusion numérique, volumes finis.

Abstract

In this work we study some numerical methods for free surface shallow water flows. On one hand, we are interested in non-hydrostatic shallow water models which take into account dispersive effects. On the other hand, we study the shallow water model with Coriolis source term and the associated geostrophic equilibrium.

First, we consider numerical methods for a family of depth-averaged Euler models from the literature. These models are shallow water models with non-hydrostatic pressure. The considered numerical methods are based on a prediction-correction method which leads to decompose the problem into two parts at each time step. The prediction part consists in solving a shallow water system which is usually solved by finite volume methods, while the correction part consists in solving an elliptic problem governing the non-hydrostatic pressure. Our first goal is to analyse the convergence of a mixed mass-lumped scheme introduced in the literature through the Gradient Discretisation Method (GDM) which is a framework comprising classic and recent discretisation schemes for diffusion problems. This method allows us to obtain a convergence result on the proposed scheme for the elliptic problem. Then, we propose a new conforming formulation of the problem on the pressure and give the corresponding error estimate provided by the GDM. Finally, an application to the $\mathbb{P}_1$ conforming finite element method is given.

Then, we aim at designing explicit colocated finite volume schemes for the nonlinear Shallow Water equations with Coriolis source term that are proved to be accurate around the geostrophic equilibrium and stable in the nonlinear framework. We define some finite volume schemes and study the two properties we are interested in : the decrease of the semi-discrete energy and the preservation of the geostrophic equilibrium in the linearized version. We also propose a conservative version of one of our scheme. Finally, we illustrate the behaviour of the schemes for some standard test cases and we exhibit a great improvement when compared to a classic finite volume scheme.

Keywords : free surface flows, shallow water equations, non-hydrostatic pressure, Gradient Discretisation Method, finite element, Coriolis force, geostrophic equilibrium, well-balanced scheme, numerical diffusion, finite volume.

Table des matières

Table des figures

Chapitre 1

Introduction

Une branche de la mécanique des fluides consiste à étudier les écoulements à surface libre, typiquement constitués d'eau plus ou moins pure, et considérés incompressibles. On entend par "surface libre", un fluide dont la surface serait contrainte uniquement par la pression atmosphérique. Historiquement, on sépare ces écoulements en deux grandes catégories : d'une part, les océans/mers qui sont considérés comme presque plats et pour lesquels la description verticale du fluide est la priorité ; d'autre part, les rivières pour lesquelles on s'intéresse le plus souvent à la hauteur d'eau et au débit. Cette distinction s'explique mathématiquement par le fait que dans le premier cas les solutions sont régulières sur de très grands domaines, alors que dans le second, il est possible d'observer des chocs *i.e.* des solutions discontinues sur des domaines réduits. On utilise donc très souvent des modèles différents en fonction de la situation rencontrée. En océanographie, on utilise des modèles de type Navier-Stokes (en pratique on néglige les effets visqueux ce qui permet d'utiliser une version simplifiée *i.e.* les équations d'Euler), alors que pour les études fluviales on préfére utiliser des modèles de type Saint-Venant. La modélisation, l'analyse et la simulation d'écoulements à surface libre sont des sujets complexes. En effet, les équations rencontrées sont à la fois difficiles à analyser et à résoudre avec de fortes non-linéarités (e.g. équations de Navier-Stokes), les échelles de temps et d'espace sont souvent très grandes et diverses (e.g. érosion des côtes), enfin, les phénomènes à modéliser sont multi-physiques et nécessitent de nombreux couplages (e.g. salinité, température).

1.1 Équations d'Euler et de Saint-Venant

Les équations d'Euler ont été introduites en 1757 par Leonhard Euler [28] dans le but de décrire des écoulements incompressibles sans viscosité et à densité uniforme égale à 1. Elles s'obtiennent à partir des équations de Navier-Stokes en négligeant les termes de friction et de viscosité. Le système d'équations a la forme suivante :

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0, \tag{1.1a}$$

$$\frac{\partial u}{\partial t} + \frac{\partial u^2}{\partial x} + \frac{\partial uv}{\partial y} + \frac{\partial uw}{\partial z} + \frac{\partial p}{\partial x} = 0, \tag{1.1b}$$

$$\frac{\partial v}{\partial t} + \frac{\partial uw}{\partial x} + \frac{\partial v^2}{\partial y} + \frac{\partial vw}{\partial z} + \frac{\partial p}{\partial y} = 0, \tag{1.1c}$$

$$\frac{\partial w}{\partial t} + \frac{\partial uw}{\partial x} + \frac{\partial vw}{\partial y} + \frac{\partial w^2}{\partial z} + \frac{\partial p}{\partial z} = g, \tag{1.1d}$$

où (u, v, w) est la vitesse, p est la pression et g est la constante de gravitation.

L'équation (1.1a) est l'équation de conservation de la masse, les équations (1.1b)-(1.1d) sont les équations de conservation de la quantité de mouvement. Numériquement ces équations sont très coûteuses à résoudre, en particulier dans le cadre de l'océanographie où les échelles de temps et d'espace sont très grandes et où les écoulements sont à surface libre.

Pour contourner cet obstacle, il est possible de réduire le système d'Euler à un modèle plus simple, celui de Saint-Venant. En effet, ce dernier est une très bonne approximation du système d'Euler pour de nombreux problèmes (rupture de barrage, inondations, charriage). Les équations dites "de Barré de Saint-Venant", publiées en 1871, décrivent les écoulements à surface libre en eaux peu profondes, d'où leur appellation anglaise *Shallow Water Equations* (SW) [46]. On les rencontre en géophysique par exemple pour décrire les écoulements de rivières, la circulation de courants océaniques, la propagation de tsunamis ou encore pour modéliser d'autres phénomènes comme les avalanches, le trafic routier ou la circulation sanguine. Du fait de sa validité expérimentale et de son efficacité numérique, le modèle SW est très utilisé pour la simulation de nombreux phénomènes d'actualité : production d'énergies renouvelables, prédiction de catastrophes naturelles, étude du climat. En effet, les deux points forts de ce modèle sont sa validité expérimentale, dépassant largement sa validité théorique, ainsi que sont excellent rapport précision / coût numérique. Nous rappelons ici brièvement son obtention à partir des équations d'Euler (1.1).

Sous l'hypothèse d'eaux peu profondes et en négligeant la pression à la surface du fluide, les équations d'Euler adimensionnées mènent à une expression de la pression $p = g(\eta - z)$ dite hydrostatique car ne dépendant que du poids du fluide à l'équilibre. Dans ce cas, on dit que la pression non-hydrostatique est négligée et le modèle SW est alors hydrostatique. Pour fermer le système, on fait de plus l'hypothèse que la vitesse est homogène selon la verticale. Ainsi en moyennant selon la verticale les équations (1.1) on obtient le modèle SW suivant :

$$\frac{\partial H}{\partial t} + \frac{\partial H\overline{u}}{\partial x} + \frac{\partial H\overline{v}}{\partial y} = 0, \tag{1.2a}$$

$$\frac{\partial H\overline{u}}{\partial t} + \frac{\partial H\overline{u}^2}{\partial x} + \frac{\partial H\overline{uv}}{\partial y} + g\frac{\partial}{\partial x}\left(\frac{H^2}{2}\right) = -gH\frac{\partial z_b}{\partial x}, \tag{1.2b}$$

$$\frac{\partial H\overline{v}}{\partial t} + \frac{\partial H\overline{uv}}{\partial x} + \frac{\partial H\overline{v}^2}{\partial y} + g\frac{\partial}{\partial y}\left(\frac{H^2}{2}\right) = -gH\frac{\partial z_b}{\partial y}, \tag{1.2c}$$

où $(\overline{u}, \overline{v})$ est la vitesse moyennée sur la verticale, H est la hauteur d'eau et z_b est la cote du fond du domaine, appelée bathymétrie. En introduisant η la cote de la surface libre on peut définir la vitesse moyenne par :

$$\overline{u} = \frac{1}{H}\int_{z_b}^{\eta} u dz, \quad \overline{v} = \frac{1}{H}\int_{z_b}^{\eta} v dz, \tag{1.3}$$

et la hauteur d'eau par :

$$H = \eta - z_b. \tag{1.4}$$

Ces différentes quantités sont illustrées dans la Figure 1.1 suivante pour le modèle SW monodimensionnel.

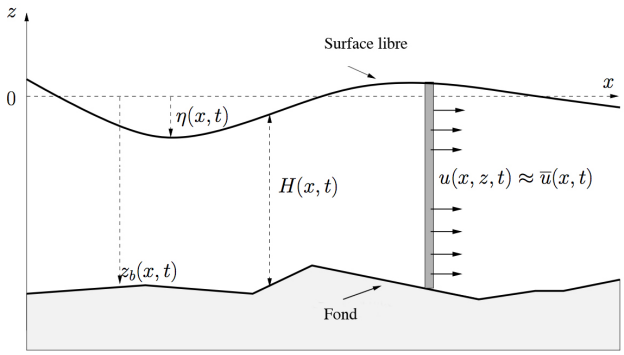


FIGURE 1.1 – Domaine et notations pour le système de Saint-Venant monodimensionnel

Par construction, la vitesse verticale w est négligeable dans les équations de Saint-Venant. Ainsi le système d'équations de Saint-Venant n'aura pas d'équation selon la direction z. La vitesse verticale w pourra néanmoins être reconstruite à partir de la solution du problème si besoin. Ce modèle permet donc d'aborder un problème physique tridimensionnel (resp. bidimensionnel) et instationnaire au travers de l'étude d'un système posé sur un domaine bidimensionnel (resp. monodimensionnel), réduisant drastiquement le coût numérique.

Cependant, tous les phénomènes physiques étudiés ne peuvent pas être modélisés de cette manière et ce modèle hydrostatique échoue à les capturer. Pour remédier à ce problème, il est possible de dériver des modèles SW non-hydrostatiques au prix d'un coût numérique et d'une complexité mathématique plus importants.

1.2 Une famille de modèles non-hydrostatiques

Le modèle SW se déduit donc des équations d'Euler via deux hypothèses majeures : une pression hydrostatique et une vitesse constante selon la verticale. On peut cependant se demander quel est le domaine de validité d'une telle approche ? Dans quel cas le modèle n'est pas suffisamment précis et doit être remplacé ? Dans la suite, on s'intéresse en particulier au cas où la pression ne peut pas être considérée hydrostatique.

Dans le cas où la pression non-hydrostatique ne peut plus être négligée, le modèle SW échoue à simuler précisément les écoulements (cf. Figure 1.3). Cette limitation a mené à l'étude de modèles dits non-hydrostatiques dans lesquels l'hypothèse de pression hydrostatique n'est pas faite. On illustre ci-dessous la différence entre un modèle SW hydrostatique et un autre non-hydrostatique. Le cas test choisi (cf. Figure 1.2) est issu du livre [21], il s'agit d'un canal, dans lequel un obstacle est présent. On génère une vague de petite amplitude sur le bord gauche du domaine et on mesure expérimentalement la hauteur d'eau en aval de l'obstacle où une condition de type libre est imposée. Puis on compare ces données aux simulations numériques pour les deux modèles. On observe le modèle hydrostatique échouer à capturer précisément les oscillations, contrairement au modèle non-hydrostatique. En effet, la présence de l'obstacle augmente la part hydrodynamique de la pression dans le fluide, causant ainsi l'apparition d'effets dispersifs non modélisés par le modèle classique SW.

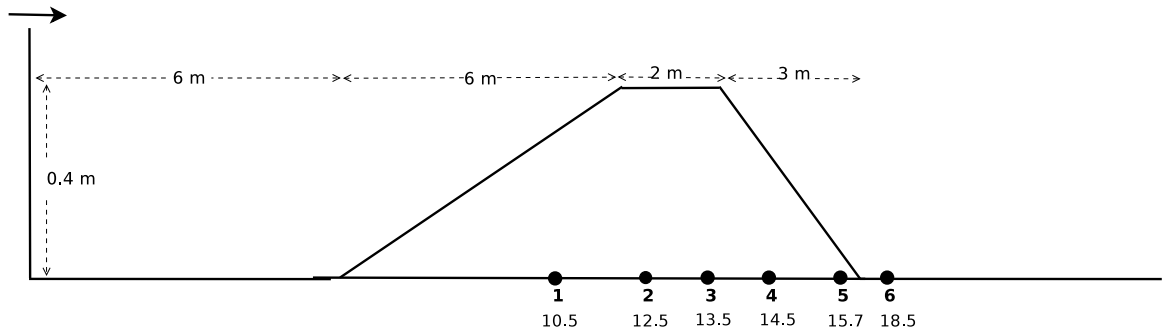


FIGURE 1.2 – Cas test de Dingemans

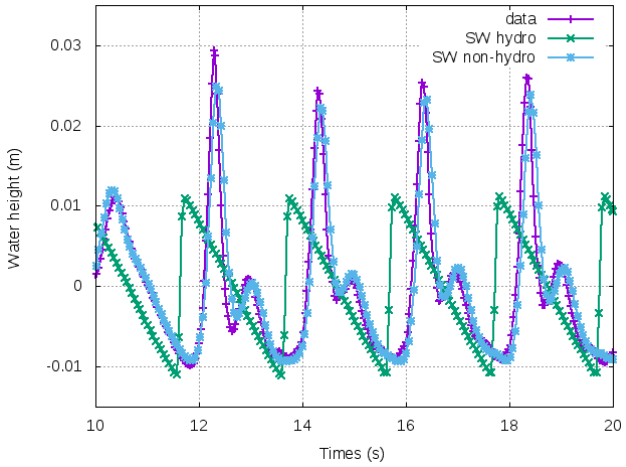


FIGURE 1.3 – Comparaison entre modèles SW hydrostatique et non-hydrostatique

Il existe de nombreux modèles non-hydrostatiques capables de rendre compte de ces effets dispersifs, chacun d'entre eux admet son propre domaine de validité. Nous nous intéressons particulièrement au modèle *depth-averaged Euler* (DAE) introduit dans [15] et étendu dans [2] en une famille de modèles dispersifs. Cette famille est issue du système d'Euler incompressible à surface libre par une approximation de type eaux peu profondes mais sans l'hypothèse hydrostatique. Le modèle est décrit par le système suivant :

$$\frac{\partial H}{\partial t} + \frac{\partial Hu}{\partial x} + \frac{\partial Hv}{\partial y} = 0, \tag{1.5a}$$

$$\frac{\partial Hu}{\partial t} + \frac{\partial}{\partial x}\left(Hu^2 + \frac{g}{2}H^2 + Hp\right) + \frac{\partial Huv}{\partial y} = -\left(gH + \frac{\alpha^2}{2}p\right)\frac{\partial z_b}{\partial x}, \tag{1.5b}$$

$$\frac{\partial Hv}{\partial t} + \frac{\partial Huv}{\partial x} + \frac{\partial}{\partial y}\left(Hv^2 + \frac{g}{2}H^2 + Hp\right) = -\left(gH + \frac{\alpha^2}{2}p\right)\frac{\partial z_b}{\partial y}, \tag{1.5c}$$

$$\frac{\partial Hw}{\partial t} + \frac{\partial Huw}{\partial x} + \frac{\partial Hvw}{\partial y} = \alpha p, \tag{1.5d}$$

$$\alpha w = -H\frac{\partial u}{\partial x} + \frac{\alpha^2}{2}u\frac{\partial z_b}{\partial x} - H\frac{\partial v}{\partial y} + \frac{\alpha^2}{2}v\frac{\partial z_b}{\partial y}, \tag{1.5e}$$

où (u, v, w) est le vecteur vitesse, p est la pression non-hydrostatique (la pression totale est donnée par

$p + gH/2$), g est la constante de gravitation et α un paramètre réel.

Ce système apparaît comme une famille de modèles SW complétés par des termes dispersifs, dépendant du paramètre α. Il peut être réécrit sous la forme condensée suivante :

$$\frac{\partial H}{\partial t} + \nabla_0 \cdot (H\boldsymbol{u}) = 0, \tag{1.6a}$$

$$\frac{\partial H\boldsymbol{u}}{\partial t} + \nabla_0 \cdot (H\boldsymbol{u} \otimes \boldsymbol{u}) + \nabla_0 \left(\frac{g}{2} H^2\right) + \nabla_s^\alpha p = -gH\nabla_0 z_b, \tag{1.6b}$$

$$\operatorname{div}_s^\alpha \boldsymbol{u} = 0, \tag{1.6c}$$

où $\nabla_0 = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, 0\right)^T$ et où les opérateurs ∇_s^α et $\operatorname{div}_s^\alpha$ sont définis par :

$$\nabla_s^\alpha \varphi = (H\nabla\varphi + \varphi\nabla\zeta, -\alpha\,\varphi)^T, \tag{1.7a}$$

$$\operatorname{div}_s^\alpha \boldsymbol{\omega} = \operatorname{div}(H\boldsymbol{v}) - \boldsymbol{v} \cdot \nabla\zeta + \alpha\, w, \tag{1.7b}$$

pour $\boldsymbol{\omega} = (\boldsymbol{v}, w)^T$ et $\zeta = H + \frac{\alpha^2}{2} z_b$.

En changeant la valeur du paramètre α on retrouve deux modèles dispersifs présent dans la littérature. En effet, $\alpha = 2$ permet d'obtenir le modèle DAE présenté dans [15], alors que $\alpha = \sqrt{3}$ permet d'écrire le modèle de Green-Naghdi [35] à un petit terme d'erreur près pour la propagation des ondes (voir [2]).

Dans [2] les auteurs proposent une stratégie numérique pour le modèle (1.6) basée sur la décomposition de Chorin-Teman appliquée au système d'Euler incompressible [44]. Par un *splitting* en temps, les auteurs obtiennent un schéma de prédiction-correction qui revient à résoudre consécutivement les deux schémas semi-discrets en temps suivants. Pour la discrétisation en temps, on introduit $t^{n+1} = t^n + \Delta t^n$, où le pas de temps Δt^n est déterminé par une condition de type CFL. L'état X^k est l'approximation de $X(t^k)$ pour $X \in \{H, \boldsymbol{u}, p\}$ et pour $k \in \{n, n+\frac{1}{2}, n+1\}$, où $t^{n+\frac{1}{2}}$ est un temps intermédiaire entre t^n et t^{n+1}.

On résout l'étape de prédiction :

$$\frac{H^{n+\frac{1}{2}} - H^n}{\Delta t^n} + \nabla_0 \cdot (H\boldsymbol{u})^n = 0, \tag{1.8a}$$

$$\frac{(H\boldsymbol{u})^{n+\frac{1}{2}} - (H\boldsymbol{u})^n}{\Delta t^n} + \nabla_0 \cdot (H\boldsymbol{u} \otimes \boldsymbol{u})^n + \nabla_0 \left(\frac{g}{2}(H^n)^2\right) = -gH^n\nabla_0 z_b. \tag{1.8b}$$

Puis, l'étape de correction :

$$H^{n+1} - H^{n+\frac{1}{2}} = 0, \tag{1.9a}$$

$$\frac{(H\boldsymbol{u})^{n+1} - (H\boldsymbol{u})^{n+\frac{1}{2}}}{\Delta t^n} + \nabla_s^\alpha p^{n+1} = 0, \tag{1.9b}$$

$$\operatorname{div}_s^\alpha \boldsymbol{u}^{n+1} = 0. \tag{1.9c}$$

Le problème (1.8) est donc le système SW (1.2) pouvant être résolu par les solveurs usuels (typiquement un schéma volumes finis). Le problème (1.9) est quant à lui un problème elliptique que les auteurs de [2] choisissent de résoudre par un schéma éléments finis mixtes. Ce schéma sera détaillé et étudié dans le Chapitre 2 de cette thèse.

1.3 Modèle de Saint-Venant avec force de Coriolis

Outre cette question sur la représentation de la pression et de la vitesse verticale à l'intérieur du fluide, le modèle SW permet la modélisation d'une grande variété de phénomènes physiques par l'ajout de termes sources correspondants. Le vent, la pluie, la topographie du fond sont autant de phénomènes pouvant être pris en compte dans le modèle SW par ce processus. Parmi ces phénomènes, on s'intéresse ici à la force de Coriolis qui résulte de la rotation de la Terre. Physiquement, son influence est importante lorsque de très grandes étendues de fluides sont considérées, typiquement de grands lacs, des mers ou des océans dans le cadre étudié ici. En océanographie, par exemple, il est essentiel de prendre en compte cette force afin d'étudier précisément la circulation des courants marins ayant un impact sur le climat. La force de Coriolis est une pseudo-force agissant perpendiculairement à la direction du mouvement d'un corps en déplacement dans un référentiel lui-même en rotation uniforme. Il ne s'agit pas d'une "force" au sens strict, *i.e.* l'action d'un corps sur un autre, mais plutôt d'une force fictive résultant du mouvement non linéaire du référentiel lui-même.

Classiquement, la force de Coriolis est représentée, elle aussi, sous la forme d'un terme source à l'équation de conservation de la quantité de mouvement du modèle SW (1.2) donné par l'expression suivante :

$$\omega H \boldsymbol{u}^{\perp}, \tag{1.10}$$

où ω est la vitesse angulaire de rotation du référentiel (ici la Terre) supposée constante, $\boldsymbol{u} = (u_x, u_y)$ est la vitesse horizontale et où $\boldsymbol{u}^{\perp} = (-u_y, u_x)$ est le vecteur orthogonal correspondant.

Considérons dans la suite un fond plat dans un souci de simplicité, le terme de topographie disparaît et le système de Saint-Venant avec force de Coriolis (SWC) s'écrit alors de manière condensée :

$$\partial_t H + \operatorname{div}(H\boldsymbol{u}) = 0\,, \tag{1.11a}$$

$$\partial_t(H\boldsymbol{u}) + \operatorname{div}(H\boldsymbol{u}\otimes\boldsymbol{u}) + H(g\nabla H + \omega\boldsymbol{u}^{\perp}) = 0\,, \tag{1.11b}$$

où g est la constante de gravitation et où H est la hauteur d'eau.

On s'intéresse maintenant aux méthodes numériques pour résoudre ce système, notre priorité est d'en assurer la stabilité et la précision.

États stationnaires : Lorsque l'on considère le système SW (1.2), *i.e.* sans le terme de Coriolis, il existe des états stationnaires non triviaux, pour lesquels les inconnues ne sont pas constantes sur le domaine. Du fait de la présence du terme source topographique, le système SW présente en effet la particularité de posséder des états stationnaires complexes. En dimension un, et pour des solutions régulières, ils sont caractérisés par les relations suivantes :

$$\frac{\partial hu}{\partial x} = 0, \quad \text{et} \quad \frac{\partial}{\partial x}\left(\frac{u^2}{2} + g(h+z_b)\right) = 0.$$

Un équilibre très important à capturer dans ce cas est "le lac au repos", défini par $u = 0$ et $h + z_b = cst.$, car de nombreuses observations expérimentales sont en réalité de petites perturbations

autour de cet équilibre. Ainsi la précision de la méthode numérique choisie est directement liée à sa capacité à préserver cet équilibre. Il est devenu usuel de construire des schémas dits équilibres (ou *well-balanced*) qui préservent le lac au repos. On citera par exemple, la reconstruction hydrostatique introduite dans [4], basée sur une discrétisation du terme source de topographie à partir d'une reconstruction de la hauteur d'eau aux interfaces des cellules. Les flux aux interfaces sont alors construits à partir de cette hauteur reconstruite. Cette technique a l'avantage d'être indépendante du solveur homogène choisi, mais adapte ce dernier au cas non-homogène en assurant la préservation de certains états stationnaires du système continu, en particulier de l'équilibre associé au lac au repos, tout en conservant intactes les propriétés de stabilité obtenues au préalable dans le cas homogène.

Cependant la présence du terme de Coriolis dans le système (1.11) amène à considérer le système linéarisé autour du "lac au repos" $(\tilde{H}, \tilde{u}) = (H_0, 0)$ avec H_0 constant :

$$\partial_t H = -H_0 \text{ div } \boldsymbol{u} \,, \tag{1.12a}$$

$$\partial_t \boldsymbol{u} = -(g\nabla H + \omega\, \boldsymbol{u}^\perp) \,. \tag{1.12b}$$

Ce système linéarisé a pour état stationnaire, l'équilibre géostrophique suivant :

$$g\nabla H + \omega \boldsymbol{u}^\perp = 0. \tag{1.13}$$

En effet, cet équilibre implique la contrainte de divergence libre :

$$\text{div } \boldsymbol{u} = 0. \tag{1.14}$$

On remarque que l'équilibre géostrophique (1.13) n'est pas un équilibre du système SWC (1.11), mais en pratique il est nécessaire que nos schémas linéarisés puissent le préserver dans un souci de stabilité. Dans [5], les auteurs ont montré que le schéma de Godunov volumes finis colocalisé classique ne préserve pas cet équilibre à cause des termes de diffusion numériques. En effet, il est nécessaire de construire des schémas dédiés dont les termes de diffusion numériques seront nuls aux alentours d'un équilibre géostrophique discret. On peut alors citer les auteurs de [14] qui ont adapté la reconstruction hydrostatique au cas avec terme source de Coriolis en introduisant artificiellement une topographie, menant alors à la méthode dite de topographie apparente (*Apparent Topography*).

Stabilité entropique : Concernant la stabilité, on remarque que le système (1.11) admet une énergie E qui vérifie (pour les solutions régulières) l'équation de conservation suivante :

$$\partial_t E + \text{div} \left[\left(gh + \frac{1}{2}\|\boldsymbol{u}\|^2 \right) h\boldsymbol{u} \right] = 0, \tag{1.15}$$

où $E = \dfrac{1}{2}gh^2 + \dfrac{1}{2}h\|\boldsymbol{u}\|^2$.

Ainsi, à l'échelle du domaine, cette énergie reste constante au cours du temps. Pour les solutions discontinues, on obtient seulement un résultat de non croissance de l'énergie. Dans tous les cas, E correspond à une entropie mathématique, en construisant une approximation numérique qui vérifie une version discrète de (1.15) on obtient alors une stabilité entropique. Les méthodes de viscosité introduites dans le cadre linéaire [5] et non-linéaire [20] fournissent un outil puissant pour assurer la stabilité des schémas au sens de l'énergie. Dans le contexte des équations de Saint-Venant non-linéaires, très peu de schémas type Volumes-Finis sont capables de fournir des inégalités d'entropie discrètes en temps et en espace, et ce même en l'absence de topographie. Dans des travaux récents [10], ces termes

de stabilisation ont été utilisés pour garantir des inégalités d'entropie discrètes pour le schéma de reconstruction hydrostatique [4], introduit pour assurer les équilibres du type lac au repos pour les équations de Saint-Venant (schémas *well-balanced*).

Dans le Chapitre 3 de cette thèse, nous proposons plusieurs schémas vérifiant ces propriétés de stabilité entropique et/ou de préservation d'équilibre géostrophique.

1.4 Organisation du manuscrit

Cette thèse se compose de deux chapitres indépendants rédigés en anglais dans le cadre de publications. Dans le Chapitre 2, réalisé en collaboration avec Cindy Guichard et Yohan Penel, nous nous intéressons au problème issu de la partie correction (1.9) du schéma de *splitting* pour le modèle DAE [2].

Soit $\Omega \subset \mathbb{R}^d$, où $d = 1$ ou $d = 2$. Le problème revient à trouver $p^\dagger \ : \ \Omega \to \mathbb{R}$ la pression non-hydrostatique et $\boldsymbol{u}^\dagger = (\boldsymbol{v}^\dagger, w^\dagger)^T \ : \ \Omega \to \mathbb{R}^{d+1}$ le champ de vitesse tels que :

$$H\boldsymbol{u}^\dagger + \nabla_s^\alpha p^\dagger = \boldsymbol{g} \text{ sur } \Omega, \tag{1.16a}$$

$$\mathrm{div}_s^\alpha \boldsymbol{u}^\dagger = f \text{ sur } \Omega, \tag{1.16b}$$

où H et $\boldsymbol{g} = (\boldsymbol{g}_1, g_2)$ sont donnés par l'étape de prédiction (1.8) et où f est la fonction nulle sur Ω dans le modèle DAE mais qui sera considérée quelconque pour l'analyse du problème ici.

En introduisant Γ_n et Γ_d tels que $\partial\Omega = \Gamma_n \cup \Gamma_d$, nous complétons le problème par les conditions de bord suivantes :

$$H\boldsymbol{u}^\dagger \cdot \boldsymbol{n}_s = H\boldsymbol{v}^\dagger \cdot \boldsymbol{n}_\Gamma = \phi \text{ sur } \Gamma_n, \tag{1.17a}$$

$$p^\dagger = 0 \text{ sur } \Gamma_d, \tag{1.17b}$$

où $\boldsymbol{n}_\Gamma$ est le vecteur normal unitaire sortant de Ω, ϕ un flux imposé et où nous définissons $\boldsymbol{n}_s := (\boldsymbol{n}_\Gamma, 0)^T$.

Nous étudions deux stratégies pour résoudre le problème (1.16)-(1.17). La première consiste à exprimer $\boldsymbol{u}^\dagger$ en fonction de $p^\dagger$ pour éliminer une inconnue en utilisant l'hypothèse théorique d'une hauteur d'eau H strictement positive. Nous pouvons alors écrire une formulation conforme en pression munie des conditions de bord correspondantes :

Trouver $p^\dagger : \Omega \to \mathbb{R}$ tel que,

$$-\mathrm{div}_s^\alpha \left(\frac{1}{H} \nabla_s^\alpha p^\dagger \right) = f - \mathrm{div}_s^\alpha \left(\frac{\boldsymbol{g}}{H} \right), \tag{1.18a}$$

$$(H\nabla p^\dagger + p^\dagger \nabla \zeta) \cdot \boldsymbol{n}_\Gamma = \boldsymbol{g}_1 \cdot \boldsymbol{n}_\Gamma - \phi \text{ sur } \Gamma_n, \tag{1.18b}$$

$$p^\dagger = 0 \text{ sur } \Gamma_d. \tag{1.18c}$$

La seconde stratégie consiste à garder la formulation mixte (1.16) munie des conditions aux limites (1.17). Les auteurs de [2] ont employé cette stratégie et ont introduit une formulation faible mixte avec un relèvement sur la vitesse, puis au niveau discret, les auteurs ont utilisé une méthode de condensation de masse (ou *mass-lumping*) pour réduire le coût numérique du schéma.

Ces deux stratégies sont étudiées dans le Chapitre 2 dans le cadre des Méthodes de Discrétisation du Gradient (GDM) [22]. Ces dernières permettent d'étudier la convergence d'une grande variété de méthodes numériques en identifiant quelques propriétés clés. Par exemple, les éléments finis conformes et non-conformes sont des GDM. Cet outil consiste à choisir des opérateurs et des espaces discrets pour la formulation faible du problème considéré, menant ainsi à un schéma numérique appelé Schéma Gradient. Dans ce chapitre, nous prouvons que le Schéma Gradient associé au problème (1.18) converge pour toute méthode appartenant aux GDM et un exemple d'application est donné pour la méthode des éléments finis conformes. Par la suite, nous montrons que le schéma éléments finis mixtes proposé dans la Section 4.1 du papier [2] est une Méthode de Discrétisation du Gradient et nous donnons les estimateurs d'erreur associés.

Dans nos travaux futurs, nous voulons implémenter le schéma éléments finis pour la formulation conforme (1.18) et le comparer au schéma proposé dans [2] en termes de précision et de temps de calcul.

Dans le Chapitre 3, réalisé en collaboration avec Emmanuel Audusse, Noémie Gaveau et Yohan Penel, nous rappelons que les schémas classiques de type Godunov pour les équations de Saint-Venant avec terme de Coriolis (1.11) manquent de précision autour de l'équilibre géostrophique (1.13). Ceci est dû au fait que les termes de diffusion qui apparaissent dans ces schémas ne s'annulent pas autour de l'équilibre géostrophique. Il est donc nécessaire de construire spécialement des schémas qui préserveront linéairement cet équilibre. Nous proposons donc des schémas impliquant une version discrétisée de l'équilibre géostrophique dans leurs termes de diffusion afin qu'ils s'annulent autour de l'équilibre. De plus, nous cherchons à avoir une stabilité entropique pour les versions semi-discrète de nos schémas lorsque c'est possible.

Nous présentons quatre schémas semi-discrets volumes finies sur maillages cartésiens qui seront comparés à un schéma HLLC de référence. Chacun de ces schémas ayant ses avantages et inconvénients, nous les présentons succinctement ici :

- le schéma *edge-based entropic* (3.7) a été présenté pour la première fois dans [25] (cf. Section A.11 de l'Annexe A) et a été reformulé dans ce chapitre dans le but d'être comparé aux schémas suivants.
  Il s'agit d'un schéma à 9 points dont les termes de diffusion sont liés à une discrétisation de l'équilibre géostrophique sur les arêtes du maillage. En revanche, l'équilibre géostrophique qui apparaît dans l'équation de conservation de quantité de mouvement (1.11b) est discrétisé aux cellules afin de s'assurer de la stabilité entropique du schéma semi-discret. Ainsi, dans ce schéma l'équilibre géostrophique est discrétisé de deux façons différentes ce qui empêche le schéma linéarisé de préserver l'une ou l'autre. Le schéma perd donc le caractère linéairement équilibre au profit de la stabilité entropique du schéma linéarisé. Cet inconvénient vient avec un avantage, contrairement aux schémas suivants, il est très difficile d'avoir une solution parasite dans le noyau du schéma linéarisé (cf. Remarque 3.4.6), en pratique nous n'en observons pas dans nos cas tests numériques en Section 3.5.

- le schéma *edge-based well-balanced* (3.14) est un schéma à 9 points dont les termes de diffusion sont liés à une discrétisation de l'équilibre géostrophique sur les arêtes du maillage. De même, l'équilibre géostrophique qui apparaît dans l'équation de conservation de quantité de mouvement (1.11b) est discrétisé sur les arêtes ce qui permet d'assurer le caractère linéairement équilibre du schéma. En revanche, la stabilité entropique du schéma semi-discret n'est plus assurée avec ce choix. En effet, un terme de création d'énergie non contrôlé apparaît. Ainsi, ce schéma est le pendant du schéma (3.7) précédent, dans lequel la stabilité entropique est sacrifiée au profit

du caractère linéairement équilibre. Autre inconvénient, des modes parasites sur la vitesse sont présents dans le noyau des schémas non-linéaire et linéaire (cf. Remarque 3.4.6).

- le schéma *cell-based* (3.16) est un schéma à 13 points dont les termes de diffusion sont liés à une discrétisation de l'équilibre géostrophique sur les cellules du maillage. De même, l'équilibre géostrophique qui apparaît dans l'équation de conservation de quantité de mouvement (1.11b) est discrétisé sur les cellules ce qui permet d'assurer le caractère linéairement équilibre du schéma. La stabilité entropique du schéma semi-discret est également assurée avec ce choix. En revanche, des modes parasites sur la hauteur d'eau sont présents dans le noyau des schémas non-linéaire et linéaire (cf. Remarque 3.4.9).
- le schéma *vertex-based* (3.20) est un schéma à 25 points dont les termes de diffusion sont liés à une discrétisation de l'équilibre géostrophique sur les nœuds du maillage. De même, l'équilibre géostrophique qui apparaît dans l'équation de conservation de quantité de mouvement (1.11b) est discrétisé sur les nœuds ce qui permet d'assurer le caractère linéairement équilibre du schéma. La stabilité entropique du schéma semi-discret est également assurée avec ce choix. En revanche, des modes parasites sur la hauteur d'eau et sur la vitesse sont présents dans le noyau des schémas non-linéaire et linéaire (cf. Remarque 3.4.15). Ce schéma admet une version conservative (3.28) qui possède les mêmes propriétés.

Concernant la discrétisation en temps, les dérivées temporelles sont discrétisées explicitement et le terme de Coriolis est discrétisé de manière semi-implicite pour tous les schémas. Ainsi, nous considérons uniquement des systèmes résolus explicitement. Le choix de la discrétisation du terme de Coriolis est justifié par les problèmes de stabilité rencontrés par le schéma HLLC dans le cas d'une discrétisation explicite (cf. Section 3.4.4). Numériquement, les autres schémas ne semblent pas nécessiter une telle discrétisation, nous l'imposons uniquement afin de comparer les schémas entre eux.

Dans la Section 3.5, nous présentons des cas tests se concentrant sur des équilibres géostrophiques afin de montrer la précision autour de ces équilibres de nos schéma par rapport à un schéma de type Godunov classique. Nous présentons également une rupture de barrage pour montrer le comportement des schémas lorsque la condition initiale est très éloignée d'un équilibre géostrophique.

Nos futurs travaux porteront sur l'étude des modes parasites présents dans les noyaux des différents schémas, le but étant de les réduire au maximum, voire de les faire disparaître. Par la suite, il sera intéressant d'étudier la stabilité des schémas entièrement discrétisés. En parallèle, la question du maillage sera explorée afin d'étendre nos travaux aux maillages triangulaires non-structurés.

Dans l'Annexe A est présenté un *proceeding* issu d'un travail collaboratif réalisé lors du CEMRACS 2019 - Geophysical Fluids, Gravity Flows. Il s'agit du travail préliminaire dans lequel le schéma *edge-based entropic* (3.7) a été introduit sous une forme différente puisque les opérateurs ont été depuis remaniés pour être cohérents avec ceux nécessaires dans les autres schémas du Chapitre 3. Une démonstration de la décroissance de l'énergie semi-discrète ainsi que le résultat de non-préservation de l'équilibre géostrophique discret peuvent être trouvés en Section A.1.1. Un schéma semi-discret sur grille décalée (A.25) est présenté en Section A.1.2. On y montre la stabilité entropique ainsi que le caractère linéairement équilibre du schéma. Enfin, en Section A.2 les deux schémas sont comparés à un schéma HLLC sur le cas test d'un vortex stationnaire et se sont révélés plus précis que ce dernier.

Chapitre 2

Gradient Discretisation Methods to analyse numerical schemes for the elliptic part of a dispersive Shallow Water system.

This work has been done in collaboration with Cindy Guichard and Yohan Penel. To be submitted

Contents

2.1 Introduction

Modelisation of free surface flows plays an important role in many engineering applications such as ocean circulation, coastal exploitation, man-made structures in rivers or lakes. Usually, this kind of flow is described by the three dimensional Navier-Stokes Equations (NSE) [32], assuming the fluid to be Newtonian, viscous and incompressible. Obviously, the real phenomena is always much more complicated, but in practice, the NSE provide a relevant mathematical model to approach the reality of oceans and river flows. At the numerical level, difficulties arise when solving these equations for real applications, especially for domains with large dimensions (e.g. oceans, lakes, rivers). Computationally, the complete resolution of the NSE for a free surface flow is known to be dramatically onerous. In order to reduce the numerical cost, we are interested in models where the dimensions are reduced. From the mathematical point of view, geophysical free surface models can be derived from the Navier-Stokes system under some assumptions.

For these reasons, and when the fluid domain can be regarded as a thin layer of fluid, it is usual to consider the nonlinear shallow water equations. This assumption implies that the model is hydrostatic and states that the velocity is constant along the water height. More generally, Non-linear Shallow Water equations (NLSW) model the dynamics of a shallow, rotating layer of homogeneous incompressible and inviscid fluid. NLSW equations are well-suited for the study and numerical simulations of a large class of geophysical phenomena, such as river flows, coastal domains, ocean circulation, or even run-off or avalanches when modified with adapted source terms. The usual conservative form of NLSW equations introduced in [46] is written as a first order hyperbolic system with various source terms such as bed slope or bottom friction terms.

However these equations rely on the hydrostatic assumption and hence when the vertical acceleration of the fluid can no longer be neglected, the shallow water system fails to represent dispersive effects e.g. in the context of wave propagation [8, 33]. It is a challenging issue to reduce a model from NSE with good properties and for which we can develop a practicable numerical method. We stress that there is not a single satisfactory model which is valid for many regimes with a good mathematical structure, each of them has a limited range of validity.

In this work, we are interested in a family of dispersive models, introduced in [2], depending on a parameter that characterises the type of model. For a given value of this parameter the authors obtain the dispersive model proposed in [1, 15] and for another value of this parameter the studied model corresponds to the Green-Naghdi model [12, 17, 35] up to some small error terms involving the bathymetry gradient. In the model considered in [2], only first order derivatives appear unlike most of existing model. The authors use a time splitting based on a Chorin-Temam projection-correction scheme [19, 47], which leads to decompose the problem into two steps at each time step. In the first one, a hyperbolic problem is solved with a finite volume scheme yielding water height H and discharge $H\boldsymbol{u}$ whereas, in the second one, an elliptic problem governing the non-hydrostatic pressure p is solved with a finite element scheme (computationally, this is the most onerous part). From a triangular primal mesh a dual mesh centered on the vertices is used for a finite volume scheme. The unknowns of the hyperbolic problem in the predictive step are constant on the dual cells. Then, for the finite element scheme, the unknowns are approximated at the vertices of the triangles. In the correction step, we use the constant values computed on the dual cell at the prediction step. In this chapter, we focus on the elliptic problem and its approximation.

Let Ω be a subset of $\mathbb{R}^d$, where $d = 1$ or $d = 2$. We introduce $\eta : \Omega \to \mathbb{R}$ the water elevation and

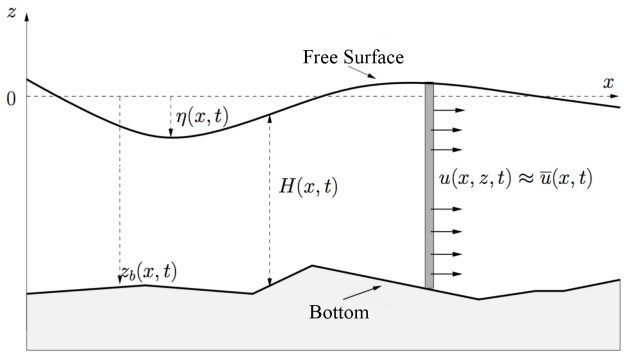


FIGURE 2.1 – Notations for the one-dimensional Shallow Water system. Source : [1]

$z_b : \Omega \to \mathbb{R}$ the bathymetry. The water height is given by $H = \eta - z_b$. Let us introduce the notation $\zeta = H + \frac{\alpha^2}{2} z_b$.

As in [2, section 2.4.1], in order to write the model under a convenient form, we introduce specific first order differential operators that we call in the following *special* gradient and *special* divergence, respectively denoted by ∇_s^α and div_s^α, which are defined for $\varphi : \Omega \to \mathbb{R}$ by,

$$\nabla_s^\alpha \varphi = (H\nabla\varphi + \varphi\nabla\zeta, -\alpha\,\varphi)^T \tag{2.1a}$$

and for $\boldsymbol{\omega} = (\boldsymbol{v}, w)^T$ with $\boldsymbol{v} : \Omega \to \mathbb{R}^d$ and $w : \Omega \to \mathbb{R}$,

$$\mathrm{div}_s^\alpha \boldsymbol{\omega} = \mathrm{div}(H\boldsymbol{v}) - \boldsymbol{v}\cdot\nabla\zeta + \alpha\, w, \tag{2.1b}$$

where, as specified above, α is the parameter which allows us to deal with several models such as the Green-Naghdi model, up to small error terms if $\alpha = \sqrt{3}$ (exact if $z_b = cst.$) [35] and the Depth-Averaged Euler model [1, 15] if $\alpha = 2$.

Then the problem reads :

Find $p^\dagger : \Omega \to \mathbb{R}$ the non-hydrostatic pressure and $\boldsymbol{u}^\dagger = (\boldsymbol{v}^\dagger, w^\dagger)^T : \Omega \to \mathbb{R}^{d+1}$ the velocity field such that,

$$H\boldsymbol{u}^\dagger + \nabla_s^\alpha p^\dagger = \boldsymbol{g} \text{ in } \Omega, \tag{2.2a}$$
$$\mathrm{div}_s^\alpha \boldsymbol{u}^\dagger = f \text{ in } \Omega, \tag{2.2b}$$

with the following boundary conditions,

$$H\boldsymbol{u}^\dagger \cdot \boldsymbol{n}_s = H\boldsymbol{v}^\dagger \cdot \boldsymbol{n}_\Gamma = \phi \text{ on } \Gamma_n, \tag{2.3}$$

and

$$p^\dagger = 0 \text{ on } \Gamma_d, \tag{2.4}$$

where $\boldsymbol{n}_\Gamma$ is the unit normal vector outward Ω and where we define $\boldsymbol{n}_s := (\boldsymbol{n}_\Gamma, 0)^T$. These boundary conditions are in the spirit of the ones used in the litterature, the reader can refer to [2, section 3.4.1] for more details about their compatibility with the full model.

We consider the following hypotheses :

- **Domain**

$$\Omega \text{ is an open connected polygonal bounded domain with Lipschitz boundary, denoted by } \Gamma, \text{ such that } \Gamma_d, \Gamma_n \text{ are two disjoint relatively open subsets of } \Gamma \text{ with } |\Gamma \backslash (\Gamma_d \cup \Gamma_n)| = 0, \quad (2.5a)$$

($|\cdot|$ denotes the $(d-1)$-dimensional measure) ;

- **Water height**

$$H \in \mathcal{C}^2(\overline{\Omega}) \text{ and there exist } \underline{H}, \overline{H} > 0 \text{ such that, for } \boldsymbol{x} \in \Omega,\ \underline{H} \leq H(\boldsymbol{x}) \leq \overline{H}, \quad (2.5b)$$

- **Other data**

$$\zeta \in \mathcal{C}^2(\overline{\Omega}),\ \boldsymbol{g} = (\boldsymbol{g}_1, g_2)^T \in L^2(\Omega)^d \times L^2(\Omega),\ \mathrm{div}(\boldsymbol{g}) \in L^2(\Omega),\ f \in L^2(\Omega),\ \phi \in L^2(\Gamma_n),\ \alpha \in \mathbb{R}^*. \quad (2.5c)$$

Our first goal is to analyse the convergence of a scheme introduced in [2] through the Gradient Discretisation Method (GDM). The GDM, detailed in [22], is a framework which contains classic and recent discretisation schemes for diffusion problems of different kinds : linear or non-linear, steady-state or time-dependent. The schemes may be conforming or non-conforming, low or high order, and may be built on very general meshes. A GDM is based on the choice of a set of discrete spaces and operators, referred to as a "Gradient Discretisation" (GD). Replacing, in the weak formulation of a diffusion problem, the continuous space and operators by the discrete elements provided by a particular GD yields a numerical scheme called a "Gradient Scheme" (GS). The concept of gradient scheme has been introduced in [31] for diffusive flows in heterogeneous anisotropic porous media with the aim of deriving a scheme which leads to symmetric positive-definite matrices. Then key properties are shown to be sufficient to prove the convergence of gradient schemes for linear and nonlinear elliptic and parabolic problems in [23]. It is shown that a number of methods are GDMs such as the conforming finite elements, the non-conforming finite elements, one discontinuous Galerkin method, the hybrid mimetic mixed or nodal mimetic finite differences methods, some discrete duality finite volume schemes, and some multi-point flux approximation schemes. The GDM has been recently extended in [24] with the Abstract Gradient Discretisation Method (AGDM), which has been developed in order to provide a unified convergence analysis that simultaneously covers all usual boundary conditions for the approximation in Banach spaces of operators acting in duality. In this work, this method allows us to study the considered problem with abstract operators and obtain a convergence result for a scheme proposed in [2] for the elliptic problem.

The chapter is organised as follows. In the next section, we propose a conforming formulation of (2.2) on the pressure and the corresponding GS. Then an application to the conforming finite element method is given. Section 2.3 is devoted to the AGDM applied to the same formulation as in Section 2.2 and to a weak mixed formulation of the strong problem (2.2) where a lifting has been operated. Finally in Section 2.4, we prove the convergence of a scheme proposed in [2] through the AGDM.

2.2 The Gradient Discretisation Method on *classic* operators

To our knowledge, Problem (2.2) has only been studied in the literature thanks to a mixed approach. We propose in this section a conforming formulation only on the pressure $p^\dagger$ which has the advantage to be easier to analyse and discretise. Then the velocity $\boldsymbol{u}^\dagger$ can be reconstructed from the pressure $p^\dagger$. We write a gradient scheme resulting from the weak problem associated to the pressure. Then, error estimates on the pressure are obtained. Finally, we choose a finite-element method as an example.

2.2.1 A conforming formulation on the pressure

From Problem (2.2), we deduce a conforming formulation by expressing the velocity $\boldsymbol{u}^\dagger$ as a function of the pressure $p^\dagger$. Indeed, from Equation (2.2a) and thanks to Assumption (2.5b) (strict positivity of H), we can express $\boldsymbol{u}^\dagger$ as follows :

$$\boldsymbol{u}^\dagger = \frac{1}{H}\left(\boldsymbol{g} - \nabla_s^\alpha p^\dagger\right) \overset{(2.1a)}{=} \frac{1}{H}\begin{pmatrix}\boldsymbol{g}_1 - H\nabla p^\dagger - p^\dagger\nabla\zeta \\ g_2 + \alpha\, p^\dagger\end{pmatrix} = \begin{pmatrix}\boldsymbol{v}^\dagger \\ w^\dagger\end{pmatrix}. \tag{2.6}$$

Remark 2.2.1

Using the classic Gradient Discretisation Method here, we deal with the classic gradient and divergence operators. In Section 2.3 we use the Abstract Gradient Discretisation framework in order to deal with the special operators (2.1a) *and* (2.1b).

From now on, the velocity $\boldsymbol{u}^\dagger$ is a secondary unknown since it can be expressed according to the main unknown $p^\dagger$. By injecting (2.6) into Equation (2.2b) we get,

$$-\mathrm{div}_s^\alpha\left(\frac{1}{H}\nabla_s^\alpha p^\dagger\right) = f - \mathrm{div}_s^\alpha\left(\frac{\boldsymbol{g}}{H}\right).$$

By using the expression of the *special* divergence (2.1b), finally the problem (2.2)-(2.4) becomes, with the *classic* operators :

Find $p^\dagger : \Omega \to \mathbb{R}$ such that,

$$-\mathrm{div}(H\nabla p^\dagger + p^\dagger\nabla\zeta) + \nabla p^\dagger\cdot\nabla\zeta + \frac{p^\dagger}{H}\left(|\nabla\zeta|^2+\alpha^2\right) = f - \mathrm{div}(\boldsymbol{g}_1) + \frac{1}{H}\left(\boldsymbol{g}_1\cdot\nabla\zeta - \alpha g_2\right) \text{ in } \Omega, \tag{2.7a}$$

$$(H\nabla p^\dagger + p^\dagger\nabla\zeta)\cdot\boldsymbol{n}_\Gamma = \boldsymbol{g}_1\cdot\boldsymbol{n}_\Gamma - \phi \text{ on } \Gamma_n, \tag{2.7b}$$

$$p^\dagger = 0 \text{ on } \Gamma_d, \tag{2.7c}$$

and with $\boldsymbol{u}^\dagger$ defined by (2.6).

2.2.2 Weak formulation

We now derive a weak formulation of Problem (2.7) in order to introduce the associated gradient scheme. Hence, a function $p^\dagger$ is said to be a weak solution of Problem (2.7) if the following holds :

Find $p^\dagger \in H_d^1(\Omega)$, such that $\forall q^\dagger \in H_d^1(\Omega)$,

$$\int_\Omega \frac{\left(H\nabla p^\dagger + p^\dagger\nabla\zeta\right)\cdot\left(H\nabla q^\dagger + q^\dagger\nabla\zeta\right)}{H} + \alpha^2\frac{p^\dagger q^\dagger}{H}\,\mathrm{d}\boldsymbol{x}$$
$$= \int_\Omega\left(f + \frac{\boldsymbol{g}_1\cdot\nabla\zeta - \alpha g_2}{H}\right)q^\dagger + \boldsymbol{g}_1\cdot\nabla q^\dagger\,\mathrm{d}\boldsymbol{x} - \int_{\Gamma_n}\phi\;\gamma(q^\dagger)\;\mathrm{d}\sigma(\boldsymbol{x}), \tag{2.8}$$

where $\gamma(\cdot)$ is the trace operator on Γ and the Hilbert space,

$$H_d^1(\Omega) = \{\varphi \in H^1(\Omega) \mid \gamma(\varphi) = 0 \text{ on } \Gamma_d\}, \tag{2.9}$$

with $\|\varphi\|^2_{H^1_d(\Omega)} = \|\varphi\|^2_{L^2(\Omega)} + \|\nabla\varphi\|^2_{L^2(\Omega)^d}$.

A solution of Problem (2.8) exists and is unique as stated in the following lemma.

Lemma 2.2.2 (*Existence and uniqueness for the conforming formulation*)

Under Assumptions (2.5)*, there exists one and only one solution to Problem* (2.8)*.*

Proof: In order to apply the Lax-Milgram theorem, we introduce the bilinear form $a(\cdot,\cdot)$ and the linear form $l(\cdot)$:

$$a(p,q) = \int_\Omega \frac{(\,H\nabla p + p\nabla\zeta\,)\cdot(\,H\nabla q + q\nabla\zeta\,)}{H} \;+\; \alpha^2\frac{pq}{H}\,\mathrm{d}\boldsymbol{x}\,,$$
$$l(q) = \int_\Omega \left(f + \frac{\boldsymbol{g}_1\cdot\nabla\zeta - \alpha g_2}{H}\right) q + \boldsymbol{g}_1\cdot\nabla q\,\mathrm{d}\boldsymbol{x} \;-\; \int_{\Gamma_n}\phi\;\gamma(q)\,\mathrm{d}\sigma(\boldsymbol{x})\;,$$

defined for all p, $q \in H^1_d(\Omega)$. We can underline that the bilinear form $a(\cdot,\cdot)$ is symmetric thus this proof could also be viewed as an application of the Riesz representation theorem. The continuity of these two forms are proved thanks to Assumptions (2.5) and Cauchy-Schwarz inequality ; the trace continuity is also needed for $l(\cdot)$. The main point is the *coercivity* – *also said ellipticity* – of $a(\cdot,\cdot)$. First of all, notice that for all ν and ω in $L^2(\Omega)^d$, we can prove that,

$$\int_\Omega \frac{|\nu|^2}{H}\mathrm{d}\boldsymbol{x} \le 2\int_\Omega \frac{|\nu+\omega|^2}{H}\mathrm{d}\boldsymbol{x} + 2\int_\Omega \frac{|\omega|^2}{H}\mathrm{d}\boldsymbol{x}\,,$$

by taking $\nu = H\nabla p^\dagger$ and $\omega = p\nabla\zeta$, we can write, also by using (2.5),

$$\underbrace{\int_\Omega H|\nabla p|^2\mathrm{d}\boldsymbol{x}}_{\ge \underline{H}\|\nabla p^\dagger\|^2_{L^2(\Omega)^d}} \le 2\underbrace{\int_\Omega \frac{|H\nabla p + p\nabla\zeta|^2}{H}\mathrm{d}\boldsymbol{x}}_{\le a(p,p)} + 2\int_\Omega \frac{p^2|\nabla\zeta|^2}{H}\mathrm{d}\boldsymbol{x}\,.$$

Since, still using (2.5),

$$\int_\Omega \frac{p^2|\nabla\zeta|^2}{H}\mathrm{d}\boldsymbol{x} \le \frac{\|\nabla\zeta\|^2_{L^\infty(\Omega)^d}}{\alpha^2}\underbrace{\int_\Omega \alpha^2\frac{p^2}{H}\mathrm{d}\boldsymbol{x}}_{\le a(p,p)}\,.$$

So we have,

$$a(p,p) \ge \frac{\alpha^2\underline{H}}{2\alpha^2 + 2\|\nabla\zeta\|^2_{L^\infty(\Omega)^d}}\|\nabla p\|^2_{L^2(\Omega)^d}\,.$$

Moreover we can write, thanks to (2.5),

$$\|p\|^2_{L^2(\Omega)} \le \frac{\overline{H}}{\alpha^2}\underbrace{\int_\Omega \frac{\alpha^2p^2}{H}\mathrm{d}\boldsymbol{x}}_{\le a(p,p)}\,.$$

Finally, we conclude the proof by writing,

$$a(p,p) \;\ge\; \frac{\alpha^2}{2}\min\left(\frac{\underline{H}}{2\alpha^2 + 2\|\nabla\zeta\|^2_{L^\infty(\Omega)^d}}\;,\;\frac{1}{\overline{H}}\right)\;\|p\|^2_{H^1_d(\Omega)}\,. \tag{2.10}$$

■

Remark 2.2.3 (*The Dirichlet boundary condition is not mandatory*)

Let us mention that the assumption $|\Gamma_d| > 0$ *is not used in order to prove Lemma 2.2.2. This is due to the term* $\alpha^2 \displaystyle\int_\Omega \frac{p^2}{H}\,\mathrm{d}\boldsymbol{x}$ *in* $a(p,p)$ *(where* $a(\cdot,\cdot)$ *is defined in the proof) which provides a bound with respect to* $\|p\|^2_{H^1_d(\Omega)}$ *without using a Poincaré inequality.*

2.2.3 Gradient Scheme

In order to introduce a gradient scheme, four discrete objects $\mathcal{D} = (X_\mathcal{D} = X_{\mathcal{D},\Omega,\Gamma_n} \oplus X_{\mathcal{D},\Gamma_d}, \Pi_\mathcal{D}, \mathbb{T}_{\mathcal{D},\Gamma_n}, \nabla_\mathcal{D})$, forming a Gradient Discretisation (GD), are introduced.

Definition 2.2.4 (*GD, homogeneous Dirichlet and non-homogeneous Neumann BCs*)

Under the assumption (2.5a)*, a gradient discretisation* $\mathcal{D}$ *for mixed boundary conditions is defined by* $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \mathbb{T}_{\mathcal{D},\Gamma_n}, \nabla_\mathcal{D})$ *where :*

1. *the set of discrete unknowns* $X_\mathcal{D} = X_{\mathcal{D},\Omega,\Gamma_n} \oplus X_{\mathcal{D},\Gamma_d}$ *is the direct sum of two finite-dimensional vector spaces on* $\mathbb{R}$*, corresponding respectively to the unknowns in* Ω *and on* Γ_n*, and with* $X_{\mathcal{D},\Gamma_d} = \{0\}$ *due to homogeneous Dirichlet condition on* Γ_d*,*
2. *the function reconstruction* $\Pi_\mathcal{D} : X_\mathcal{D} \to L^2(\Omega)$ *is linear,*
3. *the trace reconstruction* $\mathbb{T}_{\mathcal{D},\Gamma_n} : X_\mathcal{D} \to L^2(\Gamma_n)$ *is linear,*
4. *the gradient reconstruction* $\nabla_\mathcal{D} : X_\mathcal{D} \to L^2(\Omega)^d$ *is linear,*
5. *the operators are such that* $\|\cdot\|_\mathcal{D}$*, with, for* $v \in X_{\mathcal{D},\Omega,\Gamma_n}$*,* $\|v\|^2_\mathcal{D} := \|\Pi_\mathcal{D} v\|^2_{L^2(\Omega)} + \|\nabla_\mathcal{D} v\|^2_{L^2(\Omega)^d}$*, is a norm on* $X_{\mathcal{D},\Omega,\Gamma_n}$*.*

Remark 2.2.5

The definition of a GD depends on the boundary conditions (see Part I. of [22]) of the studied problem. Here, boundary conditions of the strong problem (2.7) *include a homogeneous Dirichlet condition* (2.7c) *and a Robin/Fourier condition* (2.7b)*. However in the weak formulation* (2.8)*, the term on* Γ_n *is closer to a non-homogeneous Neumann condition since there is no term under the form* $\displaystyle\int_{\Gamma_n} \gamma(p^\dagger)\,\gamma(q^\dagger)\,\mathrm{d}\sigma(\boldsymbol{x})$*. This is the reason why Definition 2.2.4 of a GD has been established for homogeneous Dirichlet and non-homogeneous Neumann conditions.*

We now introduce the Gradient Scheme (GS) for the approximation of Problem (2.8) which is designed by *replacing* the continuous spaces and operators by discrete counterparts.

Definition 2.2.6 (*GS, Conforming formulation*)

Let $\mathcal{D} = (X_{\mathcal{D}}, \Pi_{\mathcal{D}}, \mathbb{T}_{\mathcal{D},\Gamma_n}, \nabla_{\mathcal{D}})$ a GD in the sense of Definition 2.2.4. The related gradient scheme for Problem (2.8) is defined by :

Find $p \in X_{\mathcal{D},\Omega,\Gamma_n}$ such that for any $q \in X_{\mathcal{D},\Omega,\Gamma_n}$,

$$\int_\Omega \frac{(H\nabla_{\mathcal{D}} p + \Pi_{\mathcal{D}} p\ \nabla\zeta)\cdot(H\nabla_{\mathcal{D}} q + \Pi_{\mathcal{D}} q\ \nabla\zeta)}{H} + \alpha^2 \frac{\Pi_{\mathcal{D}} p\ \Pi_{\mathcal{D}} q}{H}\,\mathrm{d}\boldsymbol{x}$$
$$= \int_\Omega \left(f + \frac{\boldsymbol{g}_1\cdot\nabla\zeta - \alpha g_2}{H}\right)\Pi_{\mathcal{D}} q + \boldsymbol{g}_1\cdot\nabla_{\mathcal{D}} q\,\mathrm{d}\boldsymbol{x} - \int_{\Gamma_n} \phi\ \mathbb{T}_{\mathcal{D},\Gamma_n} q\ \mathrm{d}\sigma(\boldsymbol{x})\,. \quad (2.11)$$

2.2.4 Key properties of the considered Gradient Discretisation

In the following, we will provide general properties on the GD, given by Definition 2.2.4, that ensure the convergence of the corresponding GS (2.11). These properties are named GD-Coercivity, GD-Consistency, GD-Trace-Consistency and GD-Limit-Conformity.

Definition 2.2.7 (*GD-Coercivity*)

Under the assumption (2.5a), if $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.2.4, we define,

$$C_{\mathcal{D}} = \max_{v\in X_{\mathcal{D},\Omega,\Gamma_n}\setminus\{0\}} \left(\frac{\|\mathbb{T}_{\mathcal{D},\Gamma_n} v\|_{L^2(\Gamma_n)}}{\|v\|_{\mathcal{D}}}\right). \quad (2.12)$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.2.4 is ***coercive*** *if there exists $C_P \in \mathbb{R}_+$ such that $C_{\mathcal{D}_m} \le C_P$ for all $m \in \mathbb{N}$.*

Remark 2.2.8 (*Norm on the discrete space*)

In [22] the norm $\|\cdot\|_{\mathcal{D}}$ is mainly defined by $\|\nabla_{\mathcal{D}}\cdot\|_{L^2(\Omega)^d}$ and the GD-Coercivity introduces a kind of discrete Poincaré inequality. However, it is not the case for the problem studied here, (2.12) defines a discrete trace inequality, it is the consequence, at the discrete level, of Remark 2.2.3.

Definition 2.2.9 (*GD-Consistency*)

Under the assumption (2.5a) and using the definition (2.9) for $H^1_d(\Omega)$, if $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.2.4, we define $S_{\mathcal{D}} : H^1_d(\Omega) \to [0, +\infty)$ by,

$$\forall\varphi \in H^1_d(\Omega),\ S_{\mathcal{D}}(\varphi) = \min_{v\in X_{\mathcal{D},\Omega,\Gamma_n}} \Big\{\|\Pi_{\mathcal{D}} v - \varphi\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^2(\Omega)^d}\Big\}. \quad (2.13)$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.2.4 is ***consistent*** *if,*

$$\forall\varphi \in H^1_d(\Omega),\ \lim_{m\to\infty} S_{\mathcal{D}_m}(\varphi) = 0. \quad (2.14)$$

Definition 2.2.10 (*GD-Trace-Consistency*)

Under the assumption (2.5a) *and using the definition* (2.9) *for* $H_d^1(\Omega)$*, if* $\mathcal{D}$ *is a gradient discretisation in the sense of Definition 2.2.4, we define* $\overline{S}_{\mathcal{D}} : H_d^1(\Omega) \to [0, +\infty)$ *by,*

$$\forall \varphi \in H_d^1(\Omega),\ \overline{S}_{\mathcal{D}}(\varphi) = \min_{v \in X_{\mathcal{D},\Omega,\Gamma_n}} \left\{ \|\Pi_{\mathcal{D}} v - \varphi\|_{L^2(\Omega)} + \|\mathbb{T}_{\mathcal{D},\Gamma_n} v - \gamma(\varphi)\|_{L^2(\Gamma_n)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^2(\Omega)^d} \right\}. \tag{2.15}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ *of gradient discretisations in the sense of Definition 2.2.4 is* ***trace-consistent*** *if,*

$$\forall \varphi \in H_d^1(\Omega), \lim_{m\to\infty} \overline{S}_{\mathcal{D}_m}(\varphi) = 0. \tag{2.16}$$

The Definitions of the GD-Consistency and GD-Trace-Consistency have to be understood in the sense of an interpolation error. Moreover, we can underline that the quantity $\overline{S}_{\mathcal{D}}$ in (2.15) is needed in order to control the discrete trace $\mathbb{T}_{\mathcal{D},\Gamma_n}$.

Finally, since a GDM could be a nonconforming method, we need that the dual of the discrete gradient be "close to" a discrete divergence.

Definition 2.2.11 (*GD-Limit-Conformity*)

Under the assumption (2.5a)*, let* $\mathcal{D}$ *is a gradient discretisation in the sense of Definition 2.2.4. We introduce,*

$$H_{\mathrm{div},\Gamma_n}(\Omega) = \{\boldsymbol{\varphi} \in L^2(\Omega)^d \,:\, \mathrm{div}\boldsymbol{\varphi} \in L^2(\Omega)\ , \gamma_{\mathbf{n}}(\boldsymbol{\varphi}) \in L^2(\Gamma_n)\}, \tag{2.17}$$

where $\gamma_{\mathbf{n}}(\boldsymbol{\varphi})$ *is the normal trace of* $\boldsymbol{\varphi}$ *on* Γ_n*. We now define* $W_{\mathcal{D}} : H_{\mathrm{div},\Gamma_n}(\Omega) \to [0, +\infty)$ *by,*

$$\forall \boldsymbol{\varphi} \in H_{\mathrm{div},\Gamma_n}(\Omega),$$

$$W_{\mathcal{D}}(\boldsymbol{\varphi}) = \max_{v \in X_{\mathcal{D},\Omega,\Gamma_n}\setminus\{0\}} \frac{1}{\|v\|_{\mathcal{D}}} \left| \int_\Omega (\nabla_{\mathcal{D}} v \cdot \boldsymbol{\varphi} + \Pi_{\mathcal{D}} v\ \mathrm{div}\boldsymbol{\varphi})\ \mathrm{d}\boldsymbol{x} \ - \int_{\Gamma_n} \mathbb{T}_{\mathcal{D},\Gamma_n} v\ \gamma_{\mathbf{n}}(\boldsymbol{\varphi})\ \mathrm{d}\sigma(\boldsymbol{x}) \right|. \tag{2.18}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ *of gradient discretisations in the sense of Definition 2.2.4 is* ***limit-conforming*** *if,*

$$\forall \boldsymbol{\varphi} \in H_{\mathrm{div},\Gamma_n}(\Omega),\ \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0. \tag{2.19}$$

2.2.5 Error estimates

The error estimates of the GS (2.11) are stated in the next two theorems. We then state a corollary on the convergence result that stems from these error estimates.

Theorem 2.2.12 (*Error estimate of a GS, Conforming formulation*)

Under Assumptions (2.5), *let* $p^\dagger \in H^1_d(\Omega)$ *be the solution of Problem* (2.8). *Let* $\mathcal{D}$ *be a GD in the sense of Definition 2.2.4. Then there exists one and only one* $p \in X_{\mathcal{D},\Omega,\Gamma_n}$ *solution to the GS* (2.11) *; this solution satisfies the following inequality :*

$$\|p^\dagger - \Pi_{\mathcal{D}} p\|_{L^2(\Omega)} + \|\nabla p^\dagger - \nabla_{\mathcal{D}} p\|_{L^2(\Omega)^d} \leq S_{\mathcal{D}}(p^\dagger) + \frac{W_{\mathcal{D}}(H\nabla p^\dagger + p^\dagger \nabla\zeta - \boldsymbol{g}_1) + S_{\mathcal{D}}(p^\dagger)\, C_S}{C_{LM}}\ , \tag{2.20}$$

where $S_{\mathcal{D}}$ *and* $W_{\mathcal{D}}$ *are respectively defined by the GD-Consistency* (2.13) *and the GD-Limit-Conformity* (2.18), *moreover,* C_S *and* C_{LM} *only depend on the data of the problem (see* (2.5)*).*

Proof: Let us first prove that, if (2.20) holds for any solution $p \in X_{\mathcal{D},\Omega,\Gamma_n}$ to Scheme (2.11), then the solution to Scheme (2.8) exists and is unique. By introducing a basis of the space $X_{\mathcal{D},\Omega,\Gamma_n}$, it is easy to show that the Scheme (2.11) can be rewritten under the form of a square linear system denoted $A_{\mathcal{D}} P_{\mathcal{D}} = F_{\mathcal{D}}$ where the unknown vector $P_{\mathcal{D}}$ is made of the coordinates of p in the mentioned basis. By choosing $f = 0$, $\boldsymbol{g} = 0$ and $\phi = 0$, the right hand side of the continuous problem (2.8) vanishes and we have also $F_{\mathcal{D}} = 0$. In this case the unique solution of (2.8) is $p^\dagger = 0$. Then the right hand side of the inequality of (2.20) is equal to zero, thus we get that any solution p to the scheme satisfies $\|p\|_{\mathcal{D}} = 0$. Since $\|\cdot\|_{\mathcal{D}}$ is a norm on $X_{\mathcal{D},\Omega,\Gamma_n}$ (see Definition 2.2.4 of GD), this leads to $p = 0$. Therefore the kernel of the matrix $A_{\mathcal{D}}$ is $\{0\}$ which is equivalent to say that the linear system has a unique solution.

Let us now prove that any solution $p \in X_{\mathcal{D},\Omega,\Gamma_n}$ to Scheme (2.11) satisfies (2.20). Firstly we can notice that $p^\dagger \in H^1_d(\Omega)$ implies, in the distribution sense, that $-\mathrm{div}(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1) \in L^2(\Omega)^d$ and $(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1)\cdot\boldsymbol{n}_\Gamma = -\phi \in L^2(\Gamma_n)$ therefore, by the definition (2.17), $(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1) \in H_{\mathrm{div},\Gamma_n}(\Omega)$. Thus we can take $\boldsymbol{\varphi} = H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1$ in $W_{\mathcal{D}}$ (2.18). We then obtain, for a given $q \in X_{\mathcal{D},\Omega,\Gamma_n}$,

$$\left| \int_\Omega \nabla_{\mathcal{D}} q \cdot (H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1) + \Pi_{\mathcal{D}} q\ \mathrm{div}(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1)\ \mathrm{d}\boldsymbol{x} \right.$$
$$\left. - \int_{\Gamma_n} \mathbb{T}_{\mathcal{D},\Gamma_n} q\ \gamma_{\mathbf{n}}(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1)\,\mathrm{d}\sigma(\boldsymbol{x}) \right| \leq \|q\|_{\mathcal{D}}\, W_{\mathcal{D}}(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1).$$

Since, by (2.7a), $-\mathrm{div}(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1) = f + \dfrac{\boldsymbol{g}_1\cdot\nabla\zeta - \alpha g_2}{H} - \nabla p^\dagger\cdot\nabla\zeta - \dfrac{p^\dagger(|\nabla\zeta|^2 + \alpha^2)}{H}$ a.e. in Ω and, by (2.7b), $(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1)\cdot\boldsymbol{n}_\Gamma = -\phi$ a.e. on Γ_n, this leads to,

$$\left| \int_\Omega \nabla_{\mathcal{D}} q \cdot (H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1) - \Pi_{\mathcal{D}} q \left(f + \frac{\boldsymbol{g}_1\cdot\nabla\zeta - \alpha g_2}{H} - \nabla p^\dagger\cdot\nabla\zeta - \frac{p^\dagger(|\nabla\zeta|^2+\alpha^2)}{H} \right) \mathrm{d}\boldsymbol{x} \right.$$
$$\left. + \int_{\Gamma_n} \mathbb{T}_{\mathcal{D},\Gamma_n} q\ \phi\ \mathrm{d}\sigma(\boldsymbol{x}) \right| \leq \|q\|_{\mathcal{D}}\, W_{\mathcal{D}}(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1).$$

Now, since p is a solution of (2.11), we get,

$$\left| \int_\Omega \frac{(H\nabla_{\mathcal{D}} q + \Pi_{\mathcal{D}} q\nabla\zeta)\cdot(H(\nabla p^\dagger - \nabla_{\mathcal{D}} p) + (p^\dagger - \Pi_{\mathcal{D}} p)\nabla\zeta)}{H} + \alpha^2 \frac{\Pi_{\mathcal{D}} q(p^\dagger - \Pi_{\mathcal{D}} p)}{H}\ \mathrm{d}\boldsymbol{x} \right|$$
$$\leq \|q\|_{\mathcal{D}}\, W_{\mathcal{D}}(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1). \tag{2.21}$$

We now define,

$$I_{\mathcal{D}}p^\dagger = \underset{s\in X_{\mathcal{D},\Omega,\Gamma_n}}{\operatorname{argmin}} \left(\|\Pi_{\mathcal{D}}s - p^\dagger\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}}s - \nabla p^\dagger\|_{L^2(\Omega)^d}\right), \tag{2.22}$$

and notice that, by definition (2.13) of $S_{\mathcal{D}}$,

$$\|\Pi_{\mathcal{D}}I_{\mathcal{D}}p^\dagger - p^\dagger\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}}I_{\mathcal{D}}p^\dagger - \nabla p^\dagger\|_{L^2(\Omega)^d} = S_{\mathcal{D}}(p^\dagger). \tag{2.23}$$

We also introduce the bilinear form on $X_{\mathcal{D},\Omega,\Gamma_n} \times X_{\mathcal{D},\Omega,\Gamma_n}$,

$$a_{\mathcal{D}}(q,s) = \int_\Omega \frac{(H\nabla_{\mathcal{D}}q + \Pi_{\mathcal{D}}q\ \nabla\zeta)\cdot(H\nabla_{\mathcal{D}}s + \Pi_{\mathcal{D}}s\ \nabla\zeta)}{H} + \alpha^2\frac{\Pi_{\mathcal{D}}q\ \Pi_{\mathcal{D}}s}{H}\ \mathrm{d}\boldsymbol{x}\,.$$

By triangular inequality and by (2.21), we get,

$$\begin{aligned}\Big|a_{\mathcal{D}}(q, I_{\mathcal{D}}p^\dagger - p)\Big| \le\ & \|q\|_{\mathcal{D}}\, W_{\mathcal{D}}(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1)\\ +\Bigg|\int_\Omega & \frac{(H\nabla_{\mathcal{D}}q + \Pi_{\mathcal{D}}q\nabla\zeta)\cdot(H(\nabla_{\mathcal{D}}I_{\mathcal{D}}p^\dagger - \nabla p^\dagger) + (\Pi_{\mathcal{D}}I_{\mathcal{D}}p^\dagger - p^\dagger)\nabla\zeta)}{H} + \alpha^2\frac{\Pi_{\mathcal{D}}q\ (\Pi_{\mathcal{D}}I_{\mathcal{D}}p^\dagger - p^\dagger)}{H}\ \mathrm{d}\boldsymbol{x}\Bigg|.\end{aligned}$$

We now have to establish an upper bound of the last term of the previous inequality. Recalling the definition of $\|\cdot\|_{\mathcal{D}}$ in Definition 2.2.4, by using the Cauchy-Schwarz inequality, the assumption (2.5b) and (2.23), the right hand side is bounded by

$$\begin{aligned}\|q\|_{\mathcal{D}} \Bigg[W_{\mathcal{D}}(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1) \ +\ S_{\mathcal{D}}(p^\dagger) \underbrace{\left(\overline{H} + 2\|\nabla\zeta\|_{L^\infty(\Omega)^d} + \frac{\|\nabla\zeta\|^2_{L^\infty(\Omega)^d} + \alpha^2}{\underline{H}}\right)}_{:=\,C_S} \Bigg]\\ \ge \Big|a_{\mathcal{D}}(q, I_{\mathcal{D}}p^\dagger - p)\Big|\,.\end{aligned}$$

Choosing $q = I_{\mathcal{D}}p^\dagger - p \in X_{\mathcal{D},\Omega,\Gamma_n}$ yields,

$$0 \ \le\ a_{\mathcal{D}}(I_{\mathcal{D}}p^\dagger - p, I_{\mathcal{D}}p^\dagger - p) \ \le\ \|I_{\mathcal{D}}p^\dagger - p\|_{\mathcal{D}} \left(W_{\mathcal{D}}(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1) \ +\ S_{\mathcal{D}}(p^\dagger)\, C_S\right). \tag{2.24}$$

We now have to express a lower bound of $a_{\mathcal{D}}(I_{\mathcal{D}}p^\dagger - p, I_{\mathcal{D}}p^\dagger - p)$ with respect to $\|I_{\mathcal{D}}p^\dagger - p\|^2_{\mathcal{D}}$. To do so, it is possible to use the same tricks as in the proof of Lemma 2.2.2. And thus by mimicking at the discrete level (2.10), we obtain,

$$a_{\mathcal{D}}(I_{\mathcal{D}}p^\dagger - p, I_{\mathcal{D}}p^\dagger - p) \ \ge\ \underbrace{\frac{\alpha^2}{2}\min\left(\frac{\underline{H}}{2\alpha^2 + 2\|\nabla\zeta\|^2_{L^\infty(\Omega)^d}}\ ,\ \frac{1}{\overline{H}}\right)}_{:=\,C_{LM}>0} \|I_{\mathcal{D}}p^\dagger - p\|^2_{\mathcal{D}}\,.$$

Finally thanks to (2.24), we obtain,

$$\|I_{\mathcal{D}}p^\dagger - p\|_{\mathcal{D}} \ \le\ \frac{W_{\mathcal{D}}(H\nabla p^\dagger + p^\dagger\nabla\zeta - \boldsymbol{g}_1) + S_{\mathcal{D}}(p^\dagger)\, C_S}{C_{LM}}\,. \tag{2.25}$$

Since there holds,

$$\|p^\dagger - \Pi_\mathcal{D} p\|_{L^2(\Omega)} + \|\nabla p^\dagger - \nabla_\mathcal{D} p\|_{L^2(\Omega)^d} \leq \underbrace{\|p^\dagger - \Pi_\mathcal{D} I_\mathcal{D} p^\dagger\|_{L^2(\Omega)} + \|\nabla p^\dagger - \nabla_\mathcal{D} I_\mathcal{D} p^\dagger\|_{L^2(\Omega)^d}}_{= S_\mathcal{D}(p^\dagger)} + \|I_\mathcal{D} p^\dagger - p\|_\mathcal{D} ,$$

we obtain the estimate (2.20). ■

In [22], the obtained estimates depend on the quantity $C_\mathcal{D}$, introduced in (2.12), and related to the GD-Coercivity given by Definition 2.2.7. It is not the case here, this is due to the definition of the discrete norm $\|\cdot\|_\mathcal{D}$ which introduces a straightforward bound of $\|\Pi_\mathcal{D} \cdot\|_{L^2(\Omega)}$ with respect to $\|\cdot\|_\mathcal{D}$. The two last GD properties, i.e. GD-Coercivity and the GD-Trace-Consistency, are required to establish an estimate on the trace operator.

Proposition 2.2.13 (*Error estimate on the trace operator, Conforming formulation*)

Under Assumptions (2.5), *let* $p^\dagger \in H^1_d(\Omega)$ *be the solution of Problem* (2.8). *Let* $\mathcal{D}$ *be a GD in the sense of Definition 2.2.4 and* $p \in X_{\mathcal{D},\Omega,\Gamma_n}$ *be the solution of the GS* (2.11) *given by Theorem 2.2.12 ; this solution satisfies the following inequality :*

$$\|\gamma(p^\dagger) - \mathbb{T}_{\mathcal{D},\Gamma_n} p\|_{L^2(\Gamma_n)} \leq C_\mathcal{D} \frac{W_\mathcal{D}(H\nabla p^\dagger + p^\dagger \nabla\zeta - \boldsymbol{g}_1) + \overline{S}_\mathcal{D}(p^\dagger)\, C_S}{C_{LM}} + \overline{S}_\mathcal{D}(p^\dagger) , \tag{2.26}$$

where $C_\mathcal{D}$, $\overline{S}_\mathcal{D}$ *and* $W_\mathcal{D}$ *are respectively defined by the GD-Coercivity* (2.12), *the GD-Trace-Consistency* (2.15) *and the GD-Limit-Conformity* (2.18), *moreover,* C_S *and* C_{LM} *only depend on the data of the problem (see* (2.5)*).*

Proof: We define,

$$\overline{I}_\mathcal{D} p^\dagger = \underset{s \in X_{\mathcal{D},\Omega,\Gamma_n}}{\operatorname{argmin}} \left(\|\Pi_\mathcal{D} s - p^\dagger\|_{L^2(\Omega)} + \|\mathbb{T}_{\mathcal{D},\Gamma_n} s - \gamma(p^\dagger)\|_{L^2(\Gamma_n)} + \|\nabla_\mathcal{D} s - \nabla p^\dagger\|_{L^2(\Omega)^d} \right).$$

We can remark, by using (2.15), that

$$\|\Pi_\mathcal{D} \overline{I}_\mathcal{D} p^\dagger - p^\dagger\|_{L^2(\Omega)} + \|\mathbb{T}_{\mathcal{D},\Gamma_n} \overline{I}_\mathcal{D} p^\dagger - \gamma(p^\dagger)\|_{L^2(\Gamma_n)} + \|\nabla_\mathcal{D} \overline{I}_\mathcal{D} p^\dagger - \nabla p^\dagger\|_{L^2(\Omega)^d} = \overline{S}_\mathcal{D}(p^\dagger).$$

Moreover, by using (2.12),

$$\|\mathbb{T}_{\mathcal{D},\Gamma_n} p - \gamma(p^\dagger)\|_{L^2(\Gamma_n)} \leq \underbrace{\|\mathbb{T}_{\mathcal{D},\Gamma_n}(p - \overline{I}_\mathcal{D} p^\dagger)\|_{L^2(\Gamma_n)}}_{\leq C_\mathcal{D} \|p - \overline{I}_\mathcal{D} p^\dagger\|_\mathcal{D}} + \underbrace{\|\mathbb{T}_{\mathcal{D},\Gamma_n} \overline{I}_\mathcal{D} p^\dagger - \gamma(p^\dagger)\|_{L^2(\Gamma_n)}}_{\leq \overline{S}_\mathcal{D}(p^\dagger)}.$$

Finally, by mimicking (2.25), we can prove that,

$$\|p - \overline{I}_\mathcal{D} p^\dagger\|_\mathcal{D} \leq \frac{W_\mathcal{D}(H\nabla p^\dagger + p^\dagger \nabla\zeta - \boldsymbol{g}_1) + \overline{S}_\mathcal{D}(p^\dagger)\, C_S}{C_{LM}} ,$$

which enables us to conclude. ■

Corollary 2.2.14 (*Convergence, Conforming formulation*)

Under Assumptions (2.5)*, let* $(D_m)_{m\in\mathbb{N}}$ *be a sequence of GDs in the sense of Definition 2.2.4, which is coercive, consistent, trace-consistent and limit-conforming in the sense of Definitions 2.2.7, 2.2.9, 2.2.10 and 2.2.11.*
Then, for any $m \in \mathbb{N}$*, there exists a unique solution* $p_m \in X_{\mathcal{D}_m,\Omega,\Gamma_n}$ *to the gradient scheme* (2.11) *and, if* $p^\dagger$ *is the solution of the weak problem* (2.8) *then, as* $m \to \infty$, $\Pi_{\mathcal{D}m}p_m$ *converges to* $p^\dagger$ *in* $L^2(\Omega)$*. Similarly,* $\nabla_{\mathcal{D}m}p_m$ *converges to* $\nabla p^\dagger$ *in* $L^2(\Omega)^d$ *and* $\mathbb{T}_{\mathcal{D}m,\Gamma_n}p_m$ *converges to* $\gamma(p^\dagger)$ *in* $L^2(\Gamma_n)$*.*

Remark 2.2.15 (*Velocity reconstruction, Conforming formulation*)

Under the same assumptions and notation of the previous corollary, we can mimic the relation (2.6) *in order to define a discrete velocity* $\boldsymbol{u}_m = \begin{pmatrix} \boldsymbol{g}_1 - H\nabla_{\mathcal{D}m}p_m - \Pi_{\mathcal{D}m}p_m\nabla\zeta \\ g_2 + \alpha\,\Pi_{\mathcal{D}m}p_m \end{pmatrix}$ *which converges to* $\boldsymbol{u}^\dagger$ *in* $L^2(\Omega)^{d+1}$ *as* $m \to \infty$*.*

2.2.6 The $\mathbb{P}_1$ conforming finite elements

The $\mathbb{P}_1$ finite elements method is a particular conforming Galerkin method and thus is encompassed in the GDM framework (see section 8.1 in [22]). We briefly discuss here why the $\mathbb{P}_1$ conforming finite elements method can be seen as a GDM in the sense of Definition 2.2.4. We consider $d = 2$. Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{C}, \mathcal{V})$ be a conforming simplicial mesh of Ω in the sense of Definition 7.4 in [22] where $\mathcal{M}$ is the set of *cells*, $\mathcal{F}$ of *faces*, $\mathcal{C}$ of the *cell centers* and $\mathcal{V}$ of the *vertices*. A conforming simplicial mesh of Ω is a polytopal mesh in the sense of Definition 7.2 in [22] (see also Definition 1.50 in [27]) such that for each $K \in \mathcal{M}$ we have $Card(\mathcal{F}_K) = Card(\mathcal{V}_K) = 3$.

We introduce $N = \mathrm{card}(\mathcal{V})$ and the family $(\phi_i)_{i=1}^N$ of $\mathbb{P}_1$ finite element shape functions associated to the vertices. The $\mathbb{P}_1$ conforming finite elements method is a GDM in the sense of Definition 2.2.4 by using :

- $X_{\mathcal{D}} = X_{\mathcal{D},\Omega,\Gamma_n} \oplus X_{\mathcal{D},\Gamma_d}$ with $X_{\mathcal{D}} = \mathbb{R}^N$ and $X_{\mathcal{D},\Gamma_d} = \mathbb{R}^{N_d} = \{0\}$ where N_d is the number of vertices belonging to Γ_d,
- $\Pi_{\mathcal{D}} \ : \ X_{\mathcal{D}} \to L^2(\Omega)$ is defined, for all $\alpha := (\alpha_i)_{i=1}^N \in X_{\mathcal{D}}$, by $\Pi_{\mathcal{D}}(\alpha) = \sum_{i=1}^N \alpha_i\phi_i$ with $\alpha_i = 0$ if it is associated to a node on Γ_d ; we can remark that $\Pi_{\mathcal{D}}(\alpha) \in H_d^1(\Omega)$ which allows us to define the last two operators as follows,
- $\mathbb{T}_{\mathcal{D},\Gamma_n} \ : \ X_{\mathcal{D}} \to L^2(\Gamma_n)$ is defined by $\mathbb{T}_{\mathcal{D},\Gamma_n}(\cdot) = \gamma(\Pi_{\mathcal{D}}(\cdot))_{|\Gamma_n}$,
- $\nabla_{\mathcal{D}} \ : \ X_{\mathcal{D}} \to L^2(\Omega)^d$ is defined by $\nabla_{\mathcal{D}}(\cdot) = \nabla(\Pi_{\mathcal{D}}(\cdot))$.

Using these definitions, it is straightforward to show that $\|\cdot\|_{\mathcal{D}}$ is a norm on $X_{\mathcal{D},\Omega,\Gamma_n}$: if $\|v\|_{\mathcal{D}} = 0$ then $\Pi_{\mathcal{D}}v = 0$. Then, for all $s \in \mathcal{V}, v_s = \Pi_{\mathcal{D}}v(s) = 0$, which shows that $v = 0$.

We now list the arguments that allow us to say that, with the above definitions, the GD properties are satisfied :

- the **GD-Coercivity** is a direct application of the continuity of γ, the continuous trace operator ; this gives a uniform estimate on $C_{\mathcal{D}}$ (defined by (2.2.7)) depending only on Ω and the chosen L^p space, here L^2.

- the **GD-Consistency** and **GD-Trace-Consistency** are deduced from results on the interpolation error, see Corollary 3.23 in [27], and the continuity of γ. We can underline that an assumption on the mesh regularity is here needed as mentioned in Definition 1.107 of [27],

- the choice of discrete trace operator leads the method to be a conforming one, thus $W_{\mathcal{D}}(\cdot) = 0$ (2.18), so we have the **GD-Limit-Conformity**.

We denote by $h_{\mathcal{D}}$ the *mesh size*, i.e. the maximal diameter of the *cells*. Then, if we assume that the solution $p^\dagger$ is more regular and belongs to $H_d^2(\Omega)$, we can express the quantities $S_{\mathcal{D}}(\cdot)$ and $\overline{S}_{\mathcal{D}}(\cdot)$ as $O(h_{\mathcal{D}})$, see [27, Theorem 3.16]. This result on $\overline{S}_{\mathcal{D}}(\cdot)$ is more complex to obtain for non-conforming methods.

2.3 The Abstract Gradient Discretisation Method on *special* operators

In the previous section, we studied the strong problem (2.2)-(2.4) through the classic GDM framework based on the weak formulation of the problem with the *classic* operators ∇ and div. The Abstract Gradient Discretisation Method (AGDM) enables to mimic, at the discrete level, the *special* operators ∇_s^α and div_s^α and obtain suitable error estimates.

2.3.1 Weak formulation

First, we need to ensure that the *special* operators ∇_s^α and div_s^α defined by (2.1) fulfil a duality relation.

Lemma 2.3.1

Operators ∇_s^α and div_s^α given by (2.1) satisfy : $\forall \varphi \in H^1(\Omega)$ and $\forall \boldsymbol{\omega} \in H_{\mathrm{div}_s^\alpha}(\Omega)$,

$$\int_\Omega \nabla_s^\alpha \varphi \cdot \boldsymbol{\omega} \, \mathrm{d}\boldsymbol{x} = -\int_\Omega \varphi \, \mathrm{div}_s^\alpha \boldsymbol{\omega} \, \mathrm{d}\boldsymbol{x} + \int_\Gamma H\boldsymbol{\omega} \cdot \boldsymbol{n}_s \, \gamma(\varphi) \, \mathrm{d}\sigma(\boldsymbol{x}) \,, \tag{2.27}$$

with $H_{\mathrm{div}_s^\alpha}(\Omega) = \{\boldsymbol{\omega} \in L^2(\Omega)^{d+1} \mid \mathrm{div}_s^\alpha(\boldsymbol{\omega}) \in L^2(\Omega) \ , \ \boldsymbol{\omega} \cdot \boldsymbol{n}_s \in L^2(\Gamma) \ \}$.

Proof: Using (2.1) and Green formula we get : $\forall \varphi \in H^1(\Omega)$ and $\forall \boldsymbol{\omega} = (\boldsymbol{v}, w)^T \in H_{\mathrm{div}_s^\alpha}(\Omega)$,

$$\begin{aligned}
\int_\Omega \nabla_s^\alpha \varphi \cdot \boldsymbol{\omega} \, \mathrm{d}\boldsymbol{x} &= \int_\Omega (H\nabla\varphi + \varphi\nabla\zeta) \cdot \boldsymbol{v} - \alpha\varphi w \, \mathrm{d}\boldsymbol{x} \\
&= -\int_\Omega \varphi \, \mathrm{div}(H\boldsymbol{v}) - \varphi(\nabla\zeta \cdot \boldsymbol{v} - \alpha w) \, \mathrm{d}\boldsymbol{x} + \int_\Gamma H\boldsymbol{v} \cdot \boldsymbol{n}_\Gamma \, \gamma(\varphi) \, \mathrm{d}\sigma(\boldsymbol{x}) \\
&= -\int_\Omega \varphi \, \mathrm{div}_s^\alpha(H\boldsymbol{v}) \, \mathrm{d}\boldsymbol{x} + \int_\Gamma H\boldsymbol{\omega} \cdot \boldsymbol{n}_s \, \gamma(\varphi) \, \mathrm{d}\sigma(\boldsymbol{x}) \,.
\end{aligned}$$

■

Contrary to subsection 2.2.1, we keep the *special* operators in the conforming formulation of Problem (2.2)-(2.4). Thus, the strong problem can be read as :

Find $p^\dagger \ : \ \Omega \to \mathbb{R}$ the non hydrostatic pressure such that,

$$-\mathrm{div}_s^\alpha \left(\frac{1}{H} \nabla_s^\alpha p^\dagger \right) = f - \mathrm{div}_s^\alpha \left(\frac{1}{H} \boldsymbol{g} \right) \text{ in } \Omega, \tag{2.28a}$$

$$\nabla_s^\alpha p^\dagger \cdot \boldsymbol{n}_s = \boldsymbol{g} \cdot \boldsymbol{n}_s - \phi \text{ on } \Gamma_n, \tag{2.28b}$$

$$p^\dagger = 0 \text{ on } \Gamma_d, \tag{2.28c}$$

where we still have $\boldsymbol{u}^\dagger = \dfrac{\boldsymbol{g} - \nabla_s^\alpha p^\dagger}{H}$ as in (2.6).

We now write a weak formulation of Problem (2.28) thanks to relation (2.27) in order to introduce the associated Abstract Gradient Scheme (AGS). A function $p^\dagger$ is said to be a weak solution of Problem (2.28) if the following holds :

Find $p^\dagger \in H_d^1(\Omega)$ such that $\forall q^\dagger \in H_d^1(\Omega)$,

$$\int_\Omega \frac{\nabla_s^\alpha p^\dagger \cdot \nabla_s^\alpha q^\dagger}{H} \, \mathrm{d}\boldsymbol{x} = \int_\Omega f \ q^\dagger \, \mathrm{d}\boldsymbol{x} + \int_{\Gamma_n} -\phi \ \gamma(q^\dagger) \, \mathrm{d}\sigma(\boldsymbol{x}) + \int_\Omega \frac{\boldsymbol{g} \cdot \nabla_s^\alpha q^\dagger}{H} \, \mathrm{d}\boldsymbol{x}. \tag{2.29}$$

Equation (2.29) is Equation (2.8) using the operator ∇_s^α (2.1a). Thus existence and uniqueness of a solution are a direct consequence of Lemma 2.2.2.

2.3.2 Abstract Gradient Scheme

The approximation of Problem (2.29) relies on four discrete objects $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \mathbb{T}_{\mathcal{D},\Gamma_n}, \mathrm{G}_\mathcal{D})$ forming altogether an Abstract Gradient Discretisation.

Definition 2.3.2 (*Abstract Gradient Discretisation*)

Under the assumption (2.5a), *an abstract gradient discretisation is defined by* $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \mathbb{T}_{\mathcal{D},\Gamma_n}, \mathrm{G}_\mathcal{D})$, *where :*

1. *the set of discrete unknowns* $X_\mathcal{D} = X_{\mathcal{D},\Omega,\Gamma_n} \oplus X_{\mathcal{D},\Gamma_d}$ *is the direct sum of two finite dimensional vector spaces on* $\mathbb{R}$, *corresponding respectively to the unknowns in* Ω *and on* Γ_n, *and to the unknowns on* Γ_d ($X_{\mathcal{D},\Gamma_d} = \{0\}$ *due to* (2.4)),
2. *the function reconstruction* $\Pi_\mathcal{D} \ : \ X_\mathcal{D} \to L^2(\Omega)$ *is a linear mapping that reconstructs, from an element of* $X_\mathcal{D}$, *an element in* $L^2(\Omega)$,
3. *the trace reconstruction* $\mathbb{T}_{\mathcal{D},\Gamma_n} \ : \ X_\mathcal{D} \to L^2(\Gamma_n)$ *is a linear mapping that reconstructs, from an element of* $X_\mathcal{D}$, *an element in* Γ_n,
4. *the special gradient reconstruction* $\mathrm{G}_\mathcal{D} \ : \ X_\mathcal{D} \to L^2(\Omega)^{d+1}$ *is a linear mapping that reconstructs, from an element of* $X_\mathcal{D}$, *an element in* $L^2(\Omega)^{d+1}$,
5. *the mapping* $\mathrm{G}_\mathcal{D}$ *is such that the following quantity is a norm on* $X_{\mathcal{D},\Omega,\Gamma_n}$: $\|q\|_\mathcal{D} := \|\mathrm{G}_\mathcal{D} q\|_{L^2(\Omega)^{d+1}}$.

Now, we define the Abstract Gradient Scheme associated to the weak problem by replacing the continuous spaces and the *special* operators by discrete counterparts.

Definition 2.3.3 (*AGS, Conforming formulation*)

Let $\mathcal{D} = (X_{\mathcal{D}}, \Pi_{\mathcal{D}}, \mathbb{T}_{\mathcal{D},\Gamma_n}, \mathrm{G}_{\mathcal{D}})$ be an AGD in the sense of Definition 2.3.2. The related abstract gradient scheme for Problem (2.29) is defined by :

Find $p \in X_{\mathcal{D},\Omega,\Gamma_n}$ such that for any $q \in X_{\mathcal{D},\Omega,\Gamma_n}$,

$$\int_\Omega \frac{\mathrm{G}_{\mathcal{D}} p \cdot \mathrm{G}_{\mathcal{D}} q}{H} \,\mathrm{d}\boldsymbol{x} = \int_\Omega f\ \Pi_{\mathcal{D}} q \,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_n} -\phi\ \mathbb{T}_{\mathcal{D},\Gamma_n} q \,\mathrm{d}\sigma(\boldsymbol{x}) + \int_\Omega \frac{\boldsymbol{g} \cdot \mathrm{G}_{\mathcal{D}} q}{H} \,\mathrm{d}\boldsymbol{x}. \quad (2.30)$$

2.3.3 Key properties of the considered Abstract Gradient Discretisation

Once again, after Definition 2.3.2 of an AGD, we introduce its key properties that ensure the convergence of the corresponding AGS (2.30). These properties are defined in the following and named here AGD-Coercivity, AGD-Consistency and AGD-Limit-Conformity.

Definition 2.3.4 (*AGD-Coercivity*)

Under the assumption (2.5a), if $\mathcal{D}$ is an abstract gradient discretisation in the sense of Definition 2.3.2, let $C_{\mathcal{D}}$ be,

$$C_{\mathcal{D}} = \max_{q \in X_{\mathcal{D},\Omega,\Gamma_n} \setminus \{0\}} \frac{\|\Pi_{\mathcal{D}} q\|_{L^2(\Omega)} + \|\mathbb{T}_{\mathcal{D},\Gamma_n} q\|_{L^2(\Gamma_n)}}{\|q\|_{\mathcal{D}}}. \quad (2.31)$$

*A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of abstract gradient discretisations is **coercive** if there exists $C_P \in \mathbb{R}_+$ such that $C_{\mathcal{D}_m} \leq C_P$ for all $m \in \mathbb{N}$.*

Remark 2.3.5 (*About the definition of the coercivity*)

In comparison with the previous section, the norm $\|\cdot\|_{\mathcal{D}}$ includes a discrete version of the special gradient and thus includes the term in α. This is why, we keep here the standard definition of the coercivity regarding the literature on the GDM.

Definition 2.3.6 (*AGD-Consistency*)

Under the assumption (2.5a), if $\mathcal{D}$ is an abstract gradient discretisation in the sense of Definition 2.3.2, let $S_{\mathcal{D}} : H_d^1(\Omega) \to [0, +\infty)$ be given by,

$$\forall \varphi \in H_d^1(\Omega)\,, \quad S_{\mathcal{D}}(\varphi) = \min_{q \in X_{\mathcal{D},\Omega,\Gamma_n}} \Big(\|\Pi_{\mathcal{D}} q - \varphi\|_{L^2(\Omega)} + \|\mathbb{T}_{\mathcal{D},\Gamma_n} q - \gamma(\varphi)\|_{L^2(\Gamma_n)} + \|\mathrm{G}_{\mathcal{D}} q - \nabla_s^\alpha \varphi\|_{L^2(\Omega)^{d+1}} \Big). \quad (2.32)$$

*A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of abstract gradient discretisations is **consistent** if,*

$$\forall \varphi \in H_d^1(\Omega)\,, \ \lim_{m\to\infty} S_{\mathcal{D}_m}(\varphi) = 0. \quad (2.33)$$

The AGD-consistency can still be understood as interpolation errors but we can underline that the last term is related to the *special* gradient operator.

Definition 2.3.7 (*AGD-Limit-Conformity*)

Under the assumption (2.5a)*, if $\mathcal{D}$ is an abstract gradient discretisation in the sense of Definition 2.3.2, let $W_{\mathcal{D}} : H_{\mathrm{div}_s^\alpha}(\Omega) \to [0, +\infty)$ be given by,*

$$\forall \boldsymbol{\varphi} \in H_{\mathrm{div}_s^\alpha}(\Omega)\,,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) = \sup_{q \in X_{\mathcal{D},\Omega,\Gamma_n} \setminus \{0\}} \frac{\left| \displaystyle\int_\Omega \mathrm{G}_{\mathcal{D}} q \cdot \boldsymbol{\varphi} + \Pi_{\mathcal{D}} q \ \mathrm{div}_s^\alpha \boldsymbol{\varphi} \ \mathrm{d}\boldsymbol{x} - \int_{\Gamma_n} H \boldsymbol{\varphi} \cdot \boldsymbol{n_s} \ \mathbb{T}_{\mathcal{D},\Gamma_n} q \ \mathrm{d}\sigma(\boldsymbol{x}) \right|}{\|q\|_{\mathcal{D}}}. \tag{2.34}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of abstract gradient discretisations is ***limit-conforming*** *if,*

$$\forall \boldsymbol{\varphi} \in H_{\mathrm{div}_s^\alpha}(\Omega)\,,\ \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0.$$

The AGD-Limit-Conformity is needed in order to deal with non-conforming methods. The aim of this last property is to approach the duality relation given by Lemma (2.27).

2.3.4 Error estimates

In order to establish error estimates, we can apply Theorem 5.2 in [24] on the studied problem. The proof is close to the one done for Theorem 2.2.12 and Proposition 2.2.13 of the previous section.

Theorem 2.3.8 (*Error estimates of an AGS, Conforming formulation*)

Under Assumptions (2.5)*, let $p^\dagger \in H_d^1(\Omega)$ be the solution of Problem* (2.29)*. Let $\mathcal{D}$ be an AGD in the sense of Definition 2.3.2. Then there exists one and only one $p \in X_{\mathcal{D},\Omega,\Gamma_n}$ solution to the AGS given by* (2.30)*. This solution satisfies the following inequalities :*

$$\|\nabla_s^\alpha p^\dagger - \mathrm{G}_{\mathcal{D}} p\|_{L^2(\Omega)^{d+1}} \le \overline{H} \left[W_{\mathcal{D}} \left(\frac{\nabla_s^\alpha p^\dagger - \boldsymbol{g}}{H} \right) + \left(\frac{1}{\underline{H}} + \frac{1}{\overline{H}} \right) S_{\mathcal{D}}(p^\dagger) \right], \tag{2.35}$$

$$\|p^\dagger - \Pi_{\mathcal{D}} p\ \|_{L^2(\Omega)} + \|\gamma(p^\dagger) - \mathbb{T}_{\mathcal{D},\Gamma_n} p\|_{L^2(\Gamma_n)} \le \overline{H} \left[C_{\mathcal{D}} W_{\mathcal{D}} \left(\frac{\nabla_s^\alpha p^\dagger - \boldsymbol{g}}{H} \right) + \left(\frac{C_{\mathcal{D}}}{\underline{H}} + \frac{1}{\overline{H}} \right) S_{\mathcal{D}}(p^\dagger) \right], \tag{2.36}$$

where $C_{\mathcal{D}}$, $S_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are respectively related to the reconstruction operators, the consistency defect and the conformity defect, defined by (2.31)*,* (2.32) *and* (2.34)*.*

Proof:

Let us first prove that, if (2.35) holds for any solution $p \in X_{\mathcal{D},\Omega,\Gamma_n}$ to Scheme (2.30), then the solution exists and is unique. By introducing a basis of the space $X_{\mathcal{D},\Omega,\Gamma_n}$, it is easy to show that the Scheme (2.30) can be rewritten under the form of a square linear system denoted $A_{\mathcal{D}} P_{\mathcal{D}} = F_{\mathcal{D}}$ where the unknown vector $P_{\mathcal{D}}$ is made of the coordinates of p in the mentioned basis. By choosing $f = 0$, $\boldsymbol{g} = 0$ and $\phi = 0$ the right hand side of the continuous problem (2.29) vanishes and we also have $F_{\mathcal{D}} = 0$. In this case the unique solution of (2.29) is $p^\dagger = 0$. Then the right hand side of the inequality of (2.35) is equal to zero, thus we get that any solution p to the scheme satisfies $\|p\|_{\mathcal{D}} = 0$, which leads to $p = 0$. Therefore the kernel of the matrix $A_{\mathcal{D}}$ is $\{0\}$ which is equivalent to say that the linear system has a unique solution.

By the definition of $W_{\mathcal{D}}$ (2.34), we have $\forall q \in X_{\mathcal{D},\Omega,\Gamma_n}$ and $\forall \boldsymbol{\varphi} \in H_{\mathrm{div}_s^\alpha}(\Omega)$,

$$\left| \int_\Omega \mathrm{G}_{\mathcal{D}} q \cdot \boldsymbol{\varphi} + \Pi_{\mathcal{D}} q \ \mathrm{div}_s^\alpha \boldsymbol{\varphi} \ \mathrm{d}\boldsymbol{x} - \int_{\Gamma_n} H \boldsymbol{\varphi} \cdot \boldsymbol{n}_s \ \mathbb{T}_{\mathcal{D},\Gamma_n} q \ \mathrm{d}\sigma(\boldsymbol{x}) \right| \leq \|q\|_{\mathcal{D}} \ W_{\mathcal{D}}(\boldsymbol{\varphi}).$$

We take $\boldsymbol{\varphi} = \dfrac{1}{H}\left(\nabla_s^\alpha p^\dagger - \boldsymbol{g}\right)$ where $p^\dagger$ is the solution of (2.29). We can easily prove that $\boldsymbol{\varphi} \in H_{\mathrm{div}_s^\alpha}(\Omega)$ and, in the distribution sense, $-\mathrm{div}_s^\alpha \boldsymbol{\varphi} = f$ on Ω and $H\boldsymbol{\varphi} \cdot \boldsymbol{n}_s = -\phi$ on Γ_n. Thus we have,

$$\left| \int_\Omega \frac{1}{H} \mathrm{G}_{\mathcal{D}} q \cdot (\boldsymbol{g} - \nabla_s^\alpha p^\dagger) + \Pi_{\mathcal{D}} q \ f \ \mathrm{d}\boldsymbol{x} - \int_{\Gamma_n} \phi \ \mathbb{T}_{\mathcal{D},\Gamma_n} q \ \mathrm{d}\sigma(\boldsymbol{x}) \right| \leq \|q\|_{\mathcal{D}} \ W_{\mathcal{D}}(\boldsymbol{\varphi}).$$

Moreover, by using (2.30), we obtain,

$$\left| \int_\Omega \frac{1}{H} \mathrm{G}_{\mathcal{D}} q \cdot (\mathrm{G}_{\mathcal{D}} p - \nabla_s^\alpha p^\dagger) \ \mathrm{d}\boldsymbol{x} \right| \leq \|q\|_{\mathcal{D}} \ W_{\mathcal{D}}(\boldsymbol{\varphi}).$$

We now define,

$$I_{\mathcal{D}} p^\dagger = \underset{w \in X_{\mathcal{D},\Omega,\Gamma_n}}{\mathrm{argmin}} \left(\|\Pi_{\mathcal{D}} w - p^\dagger\|_{L^2(\Omega)} + \|\mathbb{T}_{\mathcal{D},\Gamma_n} w - \gamma(\varphi)\|_{L^2(\Gamma_n)} + \|\mathrm{G}_{\mathcal{D}} w - \nabla_s^\alpha p^\dagger\|_{L^2(\Omega)^{d+1}} \right).$$

Thus we can write that,

$$\left| \int_\Omega \frac{1}{H} \mathrm{G}_{\mathcal{D}} q \cdot (\mathrm{G}_{\mathcal{D}} p - \mathrm{G}_{\mathcal{D}} I_{\mathcal{D}} p^\dagger) \ \mathrm{d}\boldsymbol{x} \right| \leq \|q\|_{\mathcal{D}} \ W_{\mathcal{D}}(\boldsymbol{\varphi}) + \left| \int_\Omega \frac{1}{H} \mathrm{G}_{\mathcal{D}} q \cdot (\nabla_s^\alpha p^\dagger - \mathrm{G}_{\mathcal{D}} I_{\mathcal{D}} p^\dagger) \ \mathrm{d}\boldsymbol{x} \right|,$$

which implies, by the definition of $S_{\mathcal{D}}$ (2.32), and the assumption (2.5b),

$$\left| \int_\Omega \frac{1}{H} \mathrm{G}_{\mathcal{D}} q \cdot (\mathrm{G}_{\mathcal{D}} p - \mathrm{G}_{\mathcal{D}} I_{\mathcal{D}} p^\dagger) \ \mathrm{d}\boldsymbol{x} \right| \leq \|q\|_{\mathcal{D}} \left(W_{\mathcal{D}}(\boldsymbol{\varphi}) + \frac{S_{\mathcal{D}}(p^\dagger)}{\underline{H}} \right).$$

We take $q = p - I_{\mathcal{D}} p^\dagger \in X_{\mathcal{D},\Omega,\Gamma_n}$, then by using the definition of $\|\cdot\|_{\mathcal{D}}$, the Cauchy-Schwarz inequality and assumption (2.5b), we can write,

$$\frac{1}{\overline{H}} \|\mathrm{G}_{\mathcal{D}}(p - I_{\mathcal{D}} p^\dagger)\|_{L^2(\Omega)^{d+1}}^2 \leq \|\mathrm{G}_{\mathcal{D}}(p - I_{\mathcal{D}} p^\dagger)\|_{L^2(\Omega)^{d+1}} \left(W_{\mathcal{D}}(\boldsymbol{\varphi}) + \frac{S_{\mathcal{D}}(p^\dagger)}{\underline{H}} \right).$$

By using the triangular inequality and the definition of $S_{\mathcal{D}}$ (2.32), we have,

$$\|\mathrm{G}_{\mathcal{D}} p - \nabla_s^\alpha p^\dagger\|_{L^2(\Omega)^{d+1}} \leq \overline{H} \left(W_{\mathcal{D}}(\boldsymbol{\varphi}) + \frac{S_{\mathcal{D}}(p^\dagger)}{\underline{H}} \right) + S_{\mathcal{D}}(p^\dagger)$$

In order to write an estimate on the solution, we can also use Definition 2.3.4 of the AGD-Coercivity,

$$\|\Pi_{\mathcal{D}} p - p^\dagger\|_{L^2(\Omega)} + \|\mathbb{T}_{\mathcal{D},\Gamma_n} p - \gamma(p^\dagger)\|_{L^2(\Gamma_n)} \leq \overbrace{\|\Pi_{\mathcal{D}}(p - I_{\mathcal{D}} p^\dagger)\|_{L^2(\Omega)} + \|\mathbb{T}_{\mathcal{D},\Gamma_n}(p - I_{\mathcal{D}} p^\dagger)\|_{L^2(\Gamma_n)}}^{\leq C_{\mathcal{D}} \|\mathrm{G}_{\mathcal{D}}(p - I_{\mathcal{D}} p^\dagger)\|_{L^2(\Omega)^{d+1}}}$$
$$+ \underbrace{\|\Pi_{\mathcal{D}} I_{\mathcal{D}} p^\dagger - p^\dagger\|_{L^2(\Omega)} + \|\mathbb{T}_{\mathcal{D},\Gamma_n} I_{\mathcal{D}} p^\dagger - \gamma(p^\dagger)\|_{L^2(\Gamma_n)}}_{\leq S_{\mathcal{D}}(p^\dagger)}.$$

■

Remark 2.3.9

We have convergence results similar to Corollary 2.2.14 and Remark 2.2.15.

2.3.5 The $\mathbb{P}_1$ conforming finite elements

We discuss here why the $\mathbb{P}_1$ conforming finite elements method can also be seen as an Abstract Gradient Discretisation method in the sense of Definition 2.3.2. To do so we consider the same notations for the triangular mesh as in section 2.2.6. The vector of unknowns $X_{\mathcal{D}}$, the function reconstruction $\Pi_{\mathcal{D}}$ and the trace reconstruction $\mathbb{T}_{\mathcal{D},\Gamma_n}$ are also the same as in section 2.2.6. The main point is the definition of the discrete gradient.

Thus the $\mathbb{P}_1$ conforming finite elements method is a AGDM in the sense of Definition 2.3.2 by using :

- $X_{\mathcal{D}} = X_{\mathcal{D},\Omega,\Gamma_n} \oplus X_{\mathcal{D},\Gamma_d}$ with $X_{\mathcal{D}} = \mathbb{R}^N$ and $X_{\mathcal{D},\Gamma_d} = \mathbb{R}^{N_d} = \{0\}$ where N_d is the number of vertices belonging to Γ_d,
- $\Pi_{\mathcal{D}} \ : \ X_{\mathcal{D}} \to L^2(\Omega)$ is defined, for all $\alpha := (\alpha_i)_{i=1}^N \in X_{\mathcal{D}}$, by $\Pi_{\mathcal{D}}(\alpha) = \sum_{i=1}^N \alpha_i \phi_i$ with $\alpha_i = 0$ if it is associated to a node on Γ_d (ϕ_i still denote the $\mathbb{P}_1$ finite element shape function at vertex i) ; we still have $\Pi_{\mathcal{D}}(\alpha) \in H_d^1(\Omega)$ which allows us to define the last two operators as follows,
- $\mathbb{T}_{\mathcal{D},\Gamma_n} \ : \ X_{\mathcal{D}} \to L^2(\Gamma_n)$ is defined by $\mathbb{T}_{\mathcal{D},\Gamma_n}(\cdot) = \gamma(\Pi_{\mathcal{D}}(\cdot))_{|\Gamma_n}$,
- $\mathrm{G}_{\mathcal{D}} \ : \ X_{\mathcal{D}} \to L^2(\Omega)^{d+1}$ is defined by $\mathrm{G}_{\mathcal{D}}(\cdot) = \nabla_s^\alpha(\Pi_{\mathcal{D}}(\cdot)) = (H\nabla(\Pi_{\mathcal{D}}(\cdot)) + \Pi_{\mathcal{D}}(\cdot)\nabla\zeta, -\alpha\,\Pi_{\mathcal{D}}(\cdot))^T$.

Using these definitions, we have for any $v \in X_{\mathcal{D},\Omega,\Gamma_n}$,

$$\|v\|_{\mathcal{D}}^2 = \|H\nabla(\Pi_{\mathcal{D}}v) + \Pi_{\mathcal{D}}v\nabla\zeta\|_{L^2(\Omega)^d}^2 + \alpha^2\|\Pi_{\mathcal{D}}v\|_{L^2(\Omega)}^2 \ , \tag{2.37}$$

thus it is easy to show that $\|\cdot\|_{\mathcal{D}}$ is a norm on $X_{\mathcal{D},\Omega,\Gamma_n}$: if $\|v\|_{\mathcal{D}} = 0$, since $\alpha \neq 0$, then $\Pi_{\mathcal{D}}v = 0$, which implies that $v = 0$.

We now study the arguments that allow us to say that, with the above definitions, the AGD properties are satisfied :

- **AGD-Coercivity.** Thanks to (2.37) we clearly have $\|\Pi_{\mathcal{D}}v\|_{L^2(\Omega)} \leq \frac{1}{\alpha}\|v\|_{\mathcal{D}}$. Moreover we have,

  $$\|H\nabla(\Pi_{\mathcal{D}}v)\|_{L^2(\Omega)^d} \leq \|v\|_{\mathcal{D}} + \|\nabla\zeta\|_{L^\infty(\Omega)^d}\|\Pi_{\mathcal{D}}v\|_{L^2(\Omega)}.$$

  Thus, using assumption (2.5b), there exists $C > 0$ independent of the mesh, such that $\|\nabla(\Pi_{\mathcal{D}}v)\|_{L^2(\Omega)^d} \leq C\|v\|_{\mathcal{D}}$. Finally using the continuity of the trace γ, we obtain the wanted property.
- **AGD-Consistency.** This property is deduced from results on the interpolation error, see Corollary 3.23 in [27], the continuity of γ, and the assumptions on H (2.5b) and ζ (2.5c). An assumption on the mesh regularity is still needed here.
- **AGD-Limit-Conformity.** Since we deals with a conforming method, we can used the duality relation on the *special* operators, given by (2.27), to obtain $W_{\mathcal{D}}(\cdot) = 0$ (2.18).

Remark 2.3.10 (*Non-conforming construction of* $\mathrm{G}_{\mathcal{D}}$)

It is possible to make a non-conforming choice to build $\mathrm{G}_{\mathcal{D}}$. *Indeed, if we have an other function reconstruction* $\widetilde{\Pi}_{\mathcal{D}} \ : \ X_{\mathcal{D}} \to L^2(\Omega)$ *(for instance the* $\mathbb{P}_1$ *mass-lumping) we can define* $\mathrm{G}_{\mathcal{D}} \ : \ X_{\mathcal{D}} \to L^2(\Omega)^{d+1}$ *by* $\mathrm{G}_{\mathcal{D}}(\cdot) = \nabla_s^\alpha(\Pi_{\mathcal{D}}(\cdot)) = (H\nabla(\widetilde{\Pi}_{\mathcal{D}}(\cdot)) + \widetilde{\Pi}_{\mathcal{D}}(\cdot)\nabla\zeta, -\alpha\,\widetilde{\Pi}_{\mathcal{D}}(\cdot))^T$. *If the operator* $\widetilde{\Pi}_{\mathcal{D}}$ *satisfies* $\|\widetilde{\Pi}_{\mathcal{D}}(\cdot) - \Pi_{\mathcal{D}}(\cdot)\|_{L^2(\Omega)} \leq h^p\, C\, \|\mathrm{G}_{\mathcal{D}}(\cdot)\|_{L^2(\Omega)^{d+1}}$, *with* $C > 0$ *independent of the mesh and* $p > 0$, *the AGD-properties are satisfied.*

2.4 Analyse of a mixed mass-lumped scheme

As mentioned in the introduction, one goal of this work is to analyse the mixed mass-lumped schemes proposed in [2]. In this cited work, the authors introduce a mixed weak formulation with a lifting on the velocity. At the discrete level, the authors use a lumping strategy in order to reduce the numerical cost of the schemes. This section plugs these schemes into the AGD framework.

2.4.1 A weak mixed formulation with lifting

From problem (2.2)-(2.4), we write a weak mixed formulation where the functions $p^\dagger$ and $\boldsymbol{u}^\dagger$ are said to be a weak solution if the following holds :

> Find $(\boldsymbol{u}^\dagger, p^\dagger) \in H^{\phi}_{\mathrm{div}_s^\alpha}(\Omega) \times L^2(\Omega)$ such that,
>
> $$\forall \boldsymbol{\omega}^\dagger \in H^0_{\mathrm{div}_s^\alpha}(\Omega), \int_\Omega H\boldsymbol{u}^\dagger \cdot \boldsymbol{\omega}^\dagger - p^\dagger \mathrm{div}_s^\alpha \boldsymbol{\omega}^\dagger \,\mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{g}\cdot\boldsymbol{\omega}^\dagger \,\mathrm{d}\boldsymbol{x}\ , \tag{2.38a}$$
>
> $$\forall q^\dagger \in L^2(\Omega), \int_\Omega q^\dagger \mathrm{div}_s^\alpha \boldsymbol{u}^\dagger \,\mathrm{d}\boldsymbol{x} = \int_\Omega f q^\dagger \,\mathrm{d}\boldsymbol{x}\ , \tag{2.38b}$$
>
> with $H^{\phi}_{\mathrm{div}_s^\alpha}(\Omega) = \{\boldsymbol{\omega} \in L^2(\Omega)^{d+1} \mid \mathrm{div}_s^\alpha(\boldsymbol{\omega}) \in L^2(\Omega)\ ,\ H\boldsymbol{\omega}\cdot\boldsymbol{n}_s = \phi \text{ on } \Gamma_n\}$.

Now, as in [2], we introduce a lifting on the velocity $\boldsymbol{u}^\dagger$. To do so, we assume that there exist $\widetilde{\boldsymbol{\omega}} \in H^{\phi}_{\mathrm{div}_s^\alpha}(\Omega)$. Then we define $\widetilde{\boldsymbol{u}} := \boldsymbol{u}^\dagger - \widetilde{\boldsymbol{\omega}}$ and, by construction, we have $\widetilde{\boldsymbol{u}} \in H^0_{\mathrm{div}_s^\alpha}(\Omega)$. Then the problem reads :

> Find $(\widetilde{\boldsymbol{u}}, p^\dagger) \in H^0_{\mathrm{div}_s^\alpha}(\Omega) \times L^2(\Omega)$ such that,
>
> $$\forall \boldsymbol{\omega}^\dagger \in H^0_{\mathrm{div}_s^\alpha}(\Omega), \int_\Omega H\widetilde{\boldsymbol{u}} \cdot \boldsymbol{\omega}^\dagger - p^\dagger \mathrm{div}_s^\alpha \boldsymbol{\omega}^\dagger \,\mathrm{d}\boldsymbol{x} = \int_\Omega (\boldsymbol{g} - H\widetilde{\boldsymbol{\omega}})\cdot\boldsymbol{\omega}^\dagger \,\mathrm{d}\boldsymbol{x}\ , \tag{2.39a}$$
>
> $$\forall q^\dagger \in L^2(\Omega), \int_\Omega q^\dagger \mathrm{div}_s^\alpha \widetilde{\boldsymbol{u}} \,\mathrm{d}\boldsymbol{x} = \int_\Omega q^\dagger (f - \mathrm{div}_s^\alpha \widetilde{\boldsymbol{\omega}}) \,\mathrm{d}\boldsymbol{x}\ . \tag{2.39b}$$

So if we take $\widetilde{\boldsymbol{g}} := \boldsymbol{g} - H\widetilde{\boldsymbol{\omega}}$ and $\widetilde{f} := f - \mathrm{div}_s^\alpha \widetilde{\boldsymbol{\omega}}$, we recover (2.38) by dropping the tildes on f and $\boldsymbol{g}$, and by replacing $\widetilde{\boldsymbol{u}}$ with $\boldsymbol{u}^\dagger$. Thus, in the remaining of this chapter, we replace $\widetilde{f}$, respectively $\widetilde{\boldsymbol{g}}$, by f, respectively by $\boldsymbol{g}$.

We introduce as in [2], for $\boldsymbol{\omega} \in H^0_{\mathrm{div}_s^\alpha}(\Omega)$,

$$\|\boldsymbol{\omega}\|^2_{H^0_{\mathrm{div}_s^\alpha}(\Omega)} = \sum_{i=1}^{d+1} \|\boldsymbol{\omega}_i\|^2_{L^2(\Omega)} + \|\mathrm{div}_s^\alpha \boldsymbol{\omega}\|^2_{L^2(\Omega)}\ . \tag{2.40}$$

Due to the definition (2.1b), we rewrite, with $\boldsymbol{\omega} = (\boldsymbol{v}, w)^T$,

$$\|\boldsymbol{\omega}\|^2_{H^0_{\mathrm{div}_s^\alpha}(\Omega)} = \sum_{i=1}^{d} \|\boldsymbol{v}_i\|^2_{L^2(\Omega)} + \|w\|^2_{L^2(\Omega)} + \|\mathrm{div}(H\boldsymbol{v}) - \boldsymbol{v}\cdot\nabla\zeta + \alpha\, w\|^2_{L^2(\Omega)}\ .$$

The lifting on the velocity means that the boundary condition (2.3) can be considered homogeneous on Γ_n. Thus, with $\boldsymbol{u}^\dagger \in H^0_{\mathrm{div}_s^\alpha}(\Omega)$, the weak problem (2.29) involving only the pressure unknown $p^\dagger$ can be rewritten :

Find $p^\dagger \in H^1_d(\Omega)$ such that $\forall q^\dagger \in H^1_d(\Omega)$,

$$\int_\Omega \frac{\nabla_s^\alpha p^\dagger \cdot \nabla_s^\alpha q^\dagger}{H} \,\mathrm{d}\boldsymbol{x} = \int_\Omega f\ q^\dagger \,\mathrm{d}\boldsymbol{x} + \int_\Omega \frac{\boldsymbol{g} \cdot \nabla_s^\alpha q^\dagger}{H} \,\mathrm{d}\boldsymbol{x}\ . \tag{2.41}$$

The equivalence between (2.39) and (2.41) can be deduced from the *inf-sup* condition proved in [2, section 3.4.4]. The above formulation is needed in order to introduce the associated AGDM framework.

2.4.2 The associated Abstract Gradient Discretisation Method framework

In order to take into account the lifting on the velocity, we need to adapt the AGDM framework previously developed in the sections 2.3.2, 2.3.3 and 2.3.4. The adaption mainly consists in removing the terms related to Γ_n, including the trace reconstruction.

In this way, an Abstract Gradient Discretisation associated to problem (2.41) is defined as :

Definition 2.4.1 (*AGD, Lifted problem*)

Under the assumption (2.5a)*, an abstract gradient discretisation is defined by* $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \mathrm{G}_\mathcal{D})$*, where :*

1. *the set of discrete unknowns* $X_\mathcal{D} = X_{\mathcal{D},\Omega,\Gamma_n} \oplus X_{\mathcal{D},\Gamma_d}$ *is the direct sum of two finite dimensional vector spaces on* $\mathbb{R}$*, corresponding respectively to the unknowns in* Ω *and on* Γ_n*, and to the unknowns on* Γ_d *(*$X_{\mathcal{D},\Gamma_d} = \{0\}$ *due to* (2.4)*),*
2. *the function reconstruction* $\Pi_\mathcal{D} : X_\mathcal{D} \to L^2(\Omega)$ *is a linear mapping,*
3. *the special gradient reconstruction* $\mathrm{G}_\mathcal{D} : X_\mathcal{D} \to L^2(\Omega)^{d+1}$ *is a linear mapping,*
4. *the mapping* $\mathrm{G}_\mathcal{D}$ *is such that the following quantity is a norm on* $X_{\mathcal{D},\Omega,\Gamma_n}$ *:* $\|q\|_\mathcal{D} := \|\mathrm{G}_\mathcal{D} q\|_{L^2(\Omega)^{d+1}}$*.*

Now we can define the Abstract Gradient Scheme associated to the weak problem (2.41).

Definition 2.4.2 (*AGS, Lifted problem*)

Let $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \mathrm{G}_\mathcal{D})$ *an AGD in the sense of Definition 2.4.1. The related abstract gradient scheme for Problem* (2.41) *is defined by :*

Find $p \in X_{\mathcal{D},\Omega,\Gamma_n}$ *such that for any* $q \in X_{\mathcal{D},\Omega,\Gamma_n}$*,*

$$\int_\Omega \frac{\mathrm{G}_\mathcal{D} p \cdot \mathrm{G}_\mathcal{D} q}{H} \,\mathrm{d}\boldsymbol{x} = \int_\Omega f\ \Pi_\mathcal{D} q \,\mathrm{d}\boldsymbol{x} + \int_\Omega \frac{\boldsymbol{g} \cdot \mathrm{G}_\mathcal{D} q}{H} \,\mathrm{d}\boldsymbol{x}. \tag{2.42}$$

In the following, we provide the key properties of an AGD, in the sense of Definition 2.4.1, that ensure the convergence of the corresponding AGS (2.42).

Definition 2.4.3 (*AGD-Coercivity, Lifted problem*)

Under the assumption (2.5a)*, if $\mathcal{D}$ is an abstract gradient discretisation in the sense of Definition 2.4.1, let $C_{\mathcal{D}}$ be,*

$$C_{\mathcal{D}} = \max_{q \in X_{\mathcal{D},\Omega,\Gamma_n} \setminus \{0\}} \frac{\|\Pi_{\mathcal{D}} q\|_{L^2(\Omega)}}{\|q\|_{\mathcal{D}}}. \tag{2.43}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of abstract gradient discretisations is ***coercive*** *if there exists $C_P \in \mathbb{R}_+$ such that $C_{\mathcal{D}_m} \leq C_P$ for all $m \in \mathbb{N}$.*

Definition 2.4.4 (*AGD-Consistency, Lifted problem*)

Under the assumption (2.5a)*, if $\mathcal{D}$ is an abstract gradient discretisation in the sense of Definition 2.4.1, let $S_{\mathcal{D}} : H^1_d(\Omega) \to [0, +\infty)$ be given by,*

$$\forall \varphi \in H^1_d(\Omega)\,, \quad S_{\mathcal{D}}(\varphi) = \min_{q \in X_{\mathcal{D},\Omega,\Gamma_n}} \Big(\|\Pi_{\mathcal{D}} q - \varphi\|_{L^2(\Omega)} + \|\mathrm{G}_{\mathcal{D}} q - \nabla^\alpha_s \varphi\|_{L^2(\Omega)^{d+1}} \Big). \tag{2.44}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of abstract gradient discretisations is ***consistent*** *if,*

$$\forall \varphi \in H^1_d(\Omega)\,, \ \lim_{m\to\infty} S_{\mathcal{D}_m}(\varphi) = 0. \tag{2.45}$$

Definition 2.4.5 (*AGD-Limit-conformity, Lifted problem*)

Under the assumption (2.5a)*, if $\mathcal{D}$ is an abstract gradient discretisation in the sense of Definition 2.4.1, let $W_{\mathcal{D}} : H^0_{\mathrm{div}^\alpha_s}(\Omega) \to [0, +\infty)$ be given by,*

$$\forall \boldsymbol{\varphi} \in H^0_{\mathrm{div}^\alpha_s}(\Omega)\,, \ W_{\mathcal{D}}(\boldsymbol{\varphi}) = \sup_{q \in X_{\mathcal{D},\Omega,\Gamma_n} \setminus \{0\}} \frac{\left| \int_\Omega \mathrm{G}_{\mathcal{D}} q \cdot \boldsymbol{\varphi} + \Pi_{\mathcal{D}} q \ \mathrm{div}^\alpha_s \boldsymbol{\varphi} \ \mathrm{d}\boldsymbol{x} \right|}{\|q\|_{\mathcal{D}}}. \tag{2.46}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of abstract gradient discretisations is ***limit-conforming*** *if,*

$$\forall \boldsymbol{\varphi} \in H^0_{\mathrm{div}^\alpha_s}(\Omega)\,, \ \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0.$$

We now can plug Theorem 2.3.8 into the framework of Problem (2.41).

Theorem 2.4.6 (*Error estimates of an AGS, Lifted problem*)

Under Assumptions (2.5)*, let $p^\dagger \in H^1_d(\Omega)$ be the solution of Problem* (2.41)*. Let $\mathcal{D} = (X_{\mathcal{D}}, \Pi_{\mathcal{D}}, \mathrm{G}_{\mathcal{D}})$ be an AGD in the sense of Definition 2.4.1. Then there exists one and only one $p \in X_{\mathcal{D},\Omega,\Gamma_n}$ solution to the AGS* (2.42)*; this solution satisfies the following inequality :*

$$\|\Pi_{\mathcal{D}} p - p^\dagger\|_{L^2(\Omega)} \leq C_{\mathcal{D}} \overline{H} \left(W_{\mathcal{D}} \left(\frac{1}{H} \left(\nabla^\alpha_s p^\dagger - \boldsymbol{g} \right) \right) + \frac{S_{\mathcal{D}}(p^\dagger)}{\underline{H}} \right) + S_{\mathcal{D}}(p^\dagger)\,, \tag{2.47a}$$

$$\|\mathrm{G}_{\mathcal{D}} p - \nabla^\alpha_s p^\dagger\|_{L^2(\Omega)^{d+1}} \leq \overline{H} \left(W_{\mathcal{D}} \left(\frac{1}{H} \left(\nabla^\alpha_s p^\dagger - \boldsymbol{g} \right) \right) + \frac{S_{\mathcal{D}}(p^\dagger)}{\underline{H}} \right) + S_{\mathcal{D}}(p^\dagger)\,, \tag{2.47b}$$

where $S_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are respectively defined by the AGD-Consistency (2.44) *and the AGD-Limit-Conformity* (2.46)*, and C_p is related to the AGD-Coercivity* (2.43)*.*

2.4.3 Mixed mass-lumping finite element methods & AGD

In [2, section 4.] the authors applied a finite element approximation to solve the mixed problem (2.39). Thus they needed two discrete spaces, one for the velocity and one for the pressure, denoted respectively W_h^0 and Q_h in the following. The choice of these spaces is in particular done in order to ensure a discrete *inf-sup* condition. The authors used the $\mathbb{P}_1/\mathbb{P}_1$ spaces and then the $\mathbb{P}_1 - \text{iso}\mathbb{P}_2/\mathbb{P}_1$ spaces.

By the way, in both cases, the discrete approximation can be seen as choosing finite element spaces such that $W_h^0 \subset H^0_{\text{div}_s^\alpha}(\Omega)$ and $Q_h \subset L^2(\Omega)$. Then the discrete formulation of Problem (2.39) reads :

Find $(\boldsymbol{u}_h, p_h) \in W_h^0 \times Q_h$ such that,

$$\forall \boldsymbol{\omega}_h \in W_h^0, \int_\Omega H\boldsymbol{u}_h \cdot \boldsymbol{\omega}_h - p_h \ \text{div}_s^\alpha \boldsymbol{\omega}_h \ \mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{g} \cdot \boldsymbol{\omega}_h \ \mathrm{d}\boldsymbol{x} \ , \tag{2.48a}$$

$$\forall q_h \in Q_h, \int_\Omega q_h \ \text{div}_s^\alpha \boldsymbol{u}_h \ \mathrm{d}\boldsymbol{x} = \int_\Omega f q_h \ \mathrm{d}\boldsymbol{x} \ . \tag{2.48b}$$

Then the authors of [2] introduce a mass-lumping strategy in order to algebraically eliminate the velocity unknowns which leads to reduce the linear system to the discrete pressure unknowns and thus the numerical cost. After all, the authors have to resolve the following problem :

Find $(\boldsymbol{u}_h, p_h) \in W_h^0 \times Q_h$ such that,

$$\forall \boldsymbol{\omega}_h \in W_h^0, \int_\Omega H\widehat{\boldsymbol{u}}_h \cdot \widehat{\boldsymbol{\omega}}_h - p_h \ \text{div}_s^\alpha \boldsymbol{\omega}_h \ \mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{g} \cdot \widehat{\boldsymbol{\omega}}_h \ \mathrm{d}\boldsymbol{x} \ , \tag{2.49a}$$

$$\forall q_h \in Q_h, \int_\Omega q_h \ \text{div}_s^\alpha \boldsymbol{u}_h \ \mathrm{d}\boldsymbol{x} = \int_\Omega f q_h \ \mathrm{d}\boldsymbol{x} \ , \tag{2.49b}$$

where $\widehat{\cdot}_h$ is the constant piece-wise approximations on the dual cells of the finite element method.

Remark 2.4.7 (*About the mass-lumping*)

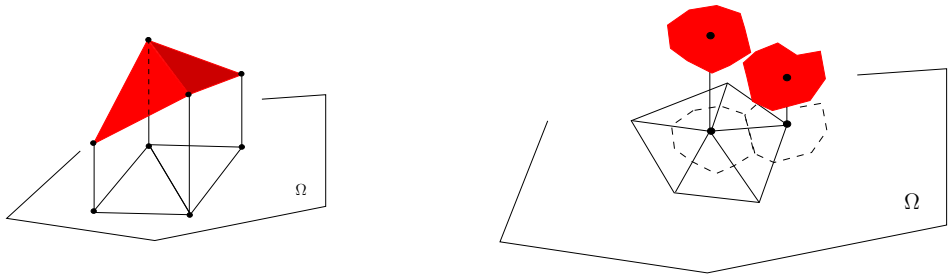
In order to better understand the notation regarding the mass-lumping, we can introduce, for $d = 2$,

$$W_h^0 = \{k \in \{1,2,3\}, u^{(k)}(\boldsymbol{x}) = \sum_{i=1}^{N} u_i^{(k)} \phi_i(\boldsymbol{x}) \ , \ (u_i^{(1)}, u_i^{(2)})^T \cdot \boldsymbol{n}_\Gamma = 0 \text{ for } \boldsymbol{x}_i \in \overline{\Gamma}_n\} \ ,$$

where N *is the number of nodes of the conforming simplicial mesh as defined in section 2.2.6, and for the the* i*-th node,* $\boldsymbol{x}_i$ *is its coordinates and* ϕ_i *the* $\mathbb{P}_1$ *finite element shape function of this vertex. Thus, if* $\boldsymbol{\omega}_h \in W_h^0$*, we can write,*

$$\boldsymbol{\omega}_h(\boldsymbol{x}) = \begin{pmatrix} \omega_h^{(1)}(\boldsymbol{x}) = \sum_{i=1}^{N} \omega_i^{(1)} \phi_i(\boldsymbol{x}) \\ \omega_h^{(2)}(\boldsymbol{x}) = \sum_{i=1}^{N} \omega_i^{(2)} \phi_i(\boldsymbol{x}) \\ \omega_h^{(3)}(\boldsymbol{x}) = \sum_{i=1}^{N} \omega_i^{(3)} \phi_i(\boldsymbol{x}) \end{pmatrix} \quad , \quad \widehat{\boldsymbol{\omega}}_h(\boldsymbol{x}) = \begin{pmatrix} \widehat{\omega}_h^{(1)}(\boldsymbol{x}) = \sum_{i=1}^{N} \omega_i^{(1)} \widehat{\phi}_i(\boldsymbol{x}) \\ \widehat{\omega}_h^{(2)}(\boldsymbol{x}) = \sum_{i=1}^{N} \omega_i^{(2)} \widehat{\phi}_i(\boldsymbol{x}) \\ \widehat{\omega}_h^{(3)}(\boldsymbol{x}) = \sum_{i=1}^{N} \omega_i^{(3)} \widehat{\phi}_i(\boldsymbol{x}) \end{pmatrix} ,$$

where $\widehat{\phi}_i$ is the piece-wise constant function equal to 1 on the dual cell K_i and 0 otherwise, where K_i corresponds to the Donald cell around the node i build using the medians of the triangles as shown at right side of the figure below.



On the figure above : at the left an example of a function $\omega_h^{(k)}$ and at the right an example of a function $\widehat{\omega}_h^{(k)}$, for $k = 1, 2$ or 3.

Remark 2.4.8 (*Discrete linear system*)

Since in [2, section 4] the mass-lumping is used at the algebrical level, the rigorous equivalence between Problem (2.49) and the schemes used in [2] should be proved through the linear system resulting of the schemes.

The aim of the end of this section is to show how the scheme (2.49) is an AGS in the sense of Definition 2.4.2. To do so, the first step is to define the appropriated AGD. We re-use the notations used in section 2.2.6 regarding the definition of a conforming simplicial mesh of Ω.

Definition 2.4.9 (*Mixed mass-lumping AGD*)

Let $\mathcal{D} = (X_{\mathcal{D}}, \Pi_{\mathcal{D}}, \mathrm{G}_{\mathcal{D}})$ be an AGD, in the sense of Definition 2.4.1, defined by :

- *$X_{\mathcal{D}} = X_{\mathcal{D},\Omega,\Gamma_n} \oplus X_{\mathcal{D},\Gamma_d}$ with $X_{\mathcal{D}} = \mathbb{R}^N$ and $X_{\mathcal{D},\Gamma_d} = \mathbb{R}^{N_d} = \{0\}$ where N_d is the number of vertices belonging to Γ_d ,*
- *$\Pi_{\mathcal{D}} : X_{\mathcal{D}} \to L^2(\Omega)$; $(p_i)_i \mapsto \sum_{i=1}^{N} p_i \varphi_i$ with $Q_h := \mathrm{span}\,(\varphi_i)_{1 \le i \le N}$,*
- *$\mathrm{G}_{\mathcal{D}} : X_{\mathcal{D}} \to L^2(\Omega)^{d+1}$; $p \mapsto \mathrm{G}_{\mathcal{D}} p$ such that :*
  - *(i) $\forall \boldsymbol{\omega}_h \in W_h^0$, $\int_\Omega \mathrm{G}_{\mathcal{D}} p \cdot \widehat{\boldsymbol{\omega}}_h + \Pi_{\mathcal{D}} p \, \mathrm{div}_s^{\alpha} \boldsymbol{\omega}_h \, \mathrm{d}\boldsymbol{x} = 0$,*
  - *(ii) $\exists G_h \in W_h^0$ such that $\mathrm{G}_{\mathcal{D}} p = H \widehat{G}_h$.*

The definition of the discrete *special* gradient $\mathrm{G}_{\mathcal{D}}$ must be specified through the following remark.

Remark 2.4.10 (*Construction of* $\mathbf{G}_{\mathcal{D}}$)

In the item (ii) we look for $G_h \in W_h^0 := \mathrm{span}\,(\psi_j)_{j\in J}$ *(where* J *is related to the d.o.f. of* W_h^0*), so we have to find* $(\beta_j)_{j\in J}$ *such that* $\widehat{G}_h = \sum_{j\in J} \beta_j \widehat{\psi}_j$*. Moreover, through the item (i), we have for any* $k \in J$ *:*

$$\int_\Omega \mathrm{G}_{\mathcal{D}} p_{\mathcal{D}} \cdot \widehat{\psi}_k \;+\; \Pi_{\mathcal{D}} p_{\mathcal{D}}\, \mathrm{div}_s^\alpha \psi_k \,\mathrm{d}\boldsymbol{x} \;=\; 0.$$

Then, by using $\mathrm{supp}(\widehat{\psi_j}) \cap \mathrm{supp}(\widehat{\psi_k}) = \emptyset$ *for* $j \neq k$ *and by denoting* $\mathrm{supp}(\widehat{\psi_k}) = K_k$*, the dual cell centered around the k-nth vertex, we obtain,*

$$\beta_k \int_{K_k} H|\psi_k|^2 \,\mathrm{d}\boldsymbol{x} = -\int_\Omega \Pi_{\mathcal{D}} p_{\mathcal{D}}\, \mathrm{div}_s^\alpha \psi_k \,\mathrm{d}\boldsymbol{x}. \tag{2.50}$$

Thus (2.50) gives a way to compute the family $(\beta_j)_{j\in J}$ *which ensures the computation of* G_h *and at last of* $\mathrm{G}_{\mathcal{D}}$*.*

Now we will state, and prove, the main result of this section.

Theorem 2.4.11 (*Equivalence between mixed mass lumping and abstract gradient scheme*)

- *On one hand, let* $(\boldsymbol{u}_h, p_h) \in W_h^0 \times Q_h$ *be the solution of the mixed mass lumping finite element scheme (2.49) with* $W_h^0 := \mathrm{span}\,(\psi_j)_{j\in J}$ *and* $Q_h := \mathrm{span}\,(\varphi_i)_{i\in I}$*.*
- *On the other hand, let* $(X_{\mathcal{D}}, \Pi_{\mathcal{D}}, \mathrm{G}_{\mathcal{D}})$ *be the AGD of Definition 2.4.9 and let* $p_{\mathcal{D}} := (\lambda_i)_{i\in I}$ *be the solution of the associated AGS (2.42).*
- *We also defined* $\boldsymbol{g}_h \in W_h^0$ *such that,*

$$\int_\Omega H\, \widehat{\boldsymbol{g}}_h \cdot \widehat{\boldsymbol{\omega}}_h \,\mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{g} \cdot \widehat{\boldsymbol{\omega}}_h \,\mathrm{d}\boldsymbol{x}\ ,\ \forall \boldsymbol{\omega}_h \in W_h^0. \tag{2.51}$$

We have the following result : $(\boldsymbol{u}_h = \boldsymbol{g}_h - G_h, p_h = \Pi_{\mathcal{D}} p_{\mathcal{D}})$ *is the solution of the mixed scheme (2.49) if and only if* $p_{\mathcal{D}}$ *is the solution of the AGS (2.42) with* $\lambda_i = \displaystyle\int_{K_i} p_h \varphi_i \,\mathrm{d}\boldsymbol{x}$ *for any vertex numbering* i*.*

Proof:

($\Rightarrow$). Let $(\boldsymbol{u}_h, p_h) \in W_h^0 \times Q_h$ the solution of scheme (2.49). We introduce $q \in X_{\mathcal{D},\Omega,\Gamma_n}$ and $q_h := \Pi_{\mathcal{D}} q \in Q_h$. We inject q_h in (2.49b). Thus we have,

$$\int_\Omega \mathrm{div}_s^\alpha \boldsymbol{u}_h\, \Pi_{\mathcal{D}} q \,\mathrm{d}\boldsymbol{x} = \int_\Omega f\ \Pi_{\mathcal{D}} q \,\mathrm{d}\boldsymbol{x}.$$

By using the item (i) in the construction of the operator $\mathrm{G}_{\mathcal{D}}$, we have,

$$\int_\Omega \mathrm{G}_{\mathcal{D}} q \cdot \widehat{\boldsymbol{u}}_h \,\mathrm{d}\boldsymbol{x} = -\int_\Omega \mathrm{div}_s^\alpha \boldsymbol{u}_h\, \Pi_{\mathcal{D}} q \,\mathrm{d}\boldsymbol{x}.$$

Thus we obtain,

$$-\int_\Omega \mathrm{G}_{\mathcal{D}} q \cdot \widehat{\boldsymbol{u}}_h \,\mathrm{d}\boldsymbol{x} = \int_\Omega f\ \Pi_{\mathcal{D}} q \,\mathrm{d}\boldsymbol{x}.$$

Now by using the item (ii) in the construction of $\mathrm{G}_{\mathcal{D}}$, there exists $F_h \in W_h^0$ such that $\mathrm{G}_{\mathcal{D}} q = H\widehat{F}_h$ with $F_h \in W_h^0$. Then by construction of $\boldsymbol{g}_h$, we can write,

$$\int_\Omega \widehat{\boldsymbol{g}}_h \cdot \mathrm{G}_{\mathcal{D}} q \, \mathrm{d}\boldsymbol{x} = \int_\Omega \frac{\boldsymbol{g} \cdot \mathrm{G}_{\mathcal{D}} q}{H} \, \mathrm{d}\boldsymbol{x} \ .$$

So we can write that,

$$\int_\Omega (\widehat{\boldsymbol{g}}_h - \widehat{\boldsymbol{u}}_h) \cdot \mathrm{G}_{\mathcal{D}} q \, \mathrm{d}\boldsymbol{x} = \int_\Omega f \ \Pi_{\mathcal{D}} q \, \mathrm{d}\boldsymbol{x} + \int_\Omega \frac{\boldsymbol{g} \cdot \mathrm{G}_{\mathcal{D}} q}{H} \, \mathrm{d}\boldsymbol{x} \ . \tag{2.52}$$

Let $p_{\mathcal{D}} := (\lambda_i)_{i \in I} \in X_{\mathcal{D},\Omega,\Gamma_n}$ such that $\lambda_i = \displaystyle\int_{K_i} p_h \varphi_i \, \mathrm{d}\boldsymbol{x}$, this choice implies that $p_h = \Pi_{\mathcal{D}} p_{\mathcal{D}}$. Thanks to the item (i) in the construction of the operator $\mathrm{G}_{\mathcal{D}}$, we can write,

$$\int_\Omega (\mathrm{G}_{\mathcal{D}} p_{\mathcal{D}} \cdot \widehat{\boldsymbol{\omega}}_h + \Pi_{\mathcal{D}} p_{\mathcal{D}} \, \mathrm{div}_s^\alpha \boldsymbol{\omega}_h) = 0 \ , \ \forall \boldsymbol{\omega}_h \in W_h^0.$$

By combining with (2.49a), we obtain,

$$\int_\Omega \mathrm{G}_{\mathcal{D}} p_{\mathcal{D}} \cdot \widehat{\boldsymbol{\omega}}_h \, \mathrm{d}\boldsymbol{x} + \int_\Omega H\widehat{\boldsymbol{u}}_h \cdot \widehat{\boldsymbol{\omega}}_h \, \mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{g} \cdot \widehat{\boldsymbol{\omega}}_h \, \mathrm{d}\boldsymbol{x} \ , \ \forall \boldsymbol{\omega}_h \in W_h^0.$$

Thanks to the construction of $\boldsymbol{\omega}_h$, we can write,

$$\int_\Omega \boldsymbol{g} \cdot \widehat{\boldsymbol{\omega}}_h \, \mathrm{d}\boldsymbol{x} = \int_\Omega H\widehat{\boldsymbol{g}}_h \cdot \widehat{\boldsymbol{\omega}}_h \, \mathrm{d}\boldsymbol{x} \ , \ \forall \boldsymbol{\omega}_h \in W_h^0.$$

At this stage, we have,

$$\int_\Omega \mathrm{G}_{\mathcal{D}} p_{\mathcal{D}} \cdot \widehat{\boldsymbol{\omega}}_h \, \mathrm{d}\boldsymbol{x} = \int_\Omega H(\widehat{\boldsymbol{g}}_h - \widehat{\boldsymbol{u}}_h) \cdot \widehat{\boldsymbol{\omega}}_h \, \mathrm{d}\boldsymbol{x} \ , \ \forall \boldsymbol{\omega}_h \in W_h^0,$$

which allows us to check, using the item (ii) of $\mathrm{G}_{\mathcal{D}}$, that $G_h = \boldsymbol{g}_h - \boldsymbol{u}_h$. To conclude, by taking $\boldsymbol{\omega}_h = \dfrac{\mathrm{G}_{\mathcal{D}} q_{\mathcal{D}}}{H}$ and by using (2.52), we recover that $p_{\mathcal{D}}$ is the solution of the AGS (2.42).

($\Leftarrow$). Let $p_{\mathcal{D}} \in X_{\mathcal{D},\Omega,\Gamma_n}$ the solution of the AGS (2.42). The item (ii) in the construction of $\mathrm{G}_{\mathcal{D}}$ implies that there exists $G_h \in W_h^0$ such that $\mathrm{G}_{\mathcal{D}} p_{\mathcal{D}} = H\widehat{G}_h$. We have defined $\boldsymbol{u}_h = \boldsymbol{g}_h - G_h$, thus $\boldsymbol{u}_h \in W_h^0$ and $\mathrm{G}_{\mathcal{D}} p_{\mathcal{D}} = H(\widehat{\boldsymbol{g}}_h - \widehat{\boldsymbol{u}}_h)$. Then the item (i) allows us to write,

$$\int_\Omega H(\widehat{\boldsymbol{g}}_h - \widehat{\boldsymbol{u}}_h) \cdot \widehat{\boldsymbol{\omega}}_h + \Pi_{\mathcal{D}} p_{\mathcal{D}} \, \mathrm{div}_s^\alpha \boldsymbol{\omega}_h \, \mathrm{d}\boldsymbol{x} \ , \ \forall \boldsymbol{\omega}_h \in W_h^0. \tag{2.53}$$

Moreover, by construction of $\boldsymbol{g}_h$, we can write,

$$\int_\Omega H\widehat{\boldsymbol{g}}_h \cdot \widehat{\boldsymbol{\omega}}_h \, \mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{g} \cdot \widehat{\boldsymbol{\omega}}_h \, \mathrm{d}\boldsymbol{x}. \tag{2.54}$$

Since we have also defined $p_h = \Pi_{\mathcal{D}} p_{\mathcal{D}}$, we have $p_h \in Q_h$ and by injecting (2.54) in (2.53), we recover the first equation (2.49a) of the mixed mass lumping finite element scheme. We now have to recover the second equation of this scheme. To do so, we inject $\mathrm{G}_{\mathcal{D}} p_{\mathcal{D}} = H(\widehat{\boldsymbol{g}}_h - \widehat{\boldsymbol{u}}_h)$ in the AGS (2.42), thus we obtain,

$$\int_\Omega (\widehat{\boldsymbol{g}}_h - \widehat{\boldsymbol{u}}_h) \cdot \mathrm{G}_{\mathcal{D}} q \, \mathrm{d}\boldsymbol{x} = \int_\Omega f \ \Pi_{\mathcal{D}} q \, \mathrm{d}\boldsymbol{x} + \int_\Omega \frac{\boldsymbol{g} \cdot \mathrm{G}_{\mathcal{D}} q}{H} \, \mathrm{d}\boldsymbol{x} \ , \ \forall q \in X_{\mathcal{D},\Omega,\Gamma_n}.$$

Thanks to the item (ii) of $\mathrm{G}_{\mathcal{D}}$, there exists $F_h \in W_h^0$ such that $\mathrm{G}_{\mathcal{D}} q = H\widehat{F}_h$ which involves, also by using the definition of $\boldsymbol{g}_h$, that,

$$\int_\Omega \widehat{\boldsymbol{g}}_h \cdot \mathrm{G}_{\mathcal{D}} q \, \mathrm{d}\boldsymbol{x} = \int_\Omega H\widehat{\boldsymbol{g}}_h \cdot \widehat{F}_h \, \mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{g} \cdot \widehat{F}_h \, \mathrm{d}\boldsymbol{x} = \int_\Omega \frac{\boldsymbol{g} \cdot \mathrm{G}_{\mathcal{D}} q}{H} \, \mathrm{d}\boldsymbol{x} \ .$$

Thus, at this stage, we have,

$$-\int_\Omega \widehat{\boldsymbol{u}}_h \cdot \mathrm{G}_{\mathcal{D}} q \, \mathrm{d}\boldsymbol{x} = \int_\Omega f \ \Pi_{\mathcal{D}} q \, \mathrm{d}\boldsymbol{x} \ , \ \forall q \in X_{\mathcal{D},\Omega,\Gamma_n}.$$

Thanks to the item (i) of $\mathrm{G}_{\mathcal{D}}$, and since $\boldsymbol{u}_h \in W_h^0$, we obtain,

$$-\int_\Omega \mathrm{G}_{\mathcal{D}} q \cdot \widehat{\boldsymbol{u}}_h \, \mathrm{d}\boldsymbol{x} = \int_\Omega \Pi_{\mathcal{D}} q \, \mathrm{div}_s^\alpha \boldsymbol{u}_h \, \mathrm{d}\boldsymbol{x} \ , \ \forall q \in X_{\mathcal{D},\Omega,\Gamma_n}.$$

So, by combining the two last equations, we have,

$$\int_\Omega \mathrm{div}_s^\alpha \boldsymbol{u}_h \Pi_{\mathcal{D}} p_{\mathcal{D}} \, \mathrm{d}\boldsymbol{x} = \int_\Omega f \ \Pi_{\mathcal{D}} q \, \mathrm{d}\boldsymbol{x} \ , \ \forall q \in X_{\mathcal{D},\Omega,\Gamma_n}.$$

Finally, by taking $q_h := \Pi_{\mathcal{D}} q$, we recover (2.49b). So we have proved that $(\boldsymbol{u}_h, p_h)$ is the solution of (2.49). ■

2.4.4 The $\mathbb{P}_1/\mathbb{P}_1$ approximation

We study here the scheme (2.49) in the case of the $\mathbb{P}_1/\mathbb{P}_1$ finite elements approximation. This method has been used in [2, subsection 4.1]. The aim of this last section is to check that, in this case, the associated AGD, see Definition 2.4.9, satisfies the mentioned AGD-properties.

As in section 2.2.6, for $d = 2$, we denote by $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{C}, \mathcal{V})$ a conforming simplicial mesh of Ω in the sense of Definition 7.4 in [22] where $\mathcal{M}$ is the set of *cells* (or *triangles*), $\mathcal{F}$ of *faces* (or *edges*), $\mathcal{C}$ of the *cell centers* and $\mathcal{V}$ of the *vertices* (or *nodes*). A conforming simplicial mesh of Ω is a polytopal mesh in the sense of Definition 7.2 in [22] (see also Definition 1.50 in [27]) such that for each $K \in \mathcal{M}$ we have $Card(\mathcal{F}_K) = Card(\mathcal{V}_K) = 3$. We then introduce $N = \mathrm{card}(\mathcal{V})$ and the family $(\phi_i)_{i=1}^N$ of $\mathbb{P}_1$ finite element shape functions associated to the vertices. We can now introduce the following space,

$$V_h = \{v_h \in \mathcal{C}^0(\overline{\Omega}) \ , \ v_{h|K} \in \mathbb{P}_1(\mathbb{R}^d) \ , \ \forall K \in \mathcal{M}\}.$$

So we take, in the AGD-Defintion 2.4.9,$Q_h = V_h$ and $W_h^0 = \{\boldsymbol{u}_h \in V_h^{d+1} \ , \ \boldsymbol{u}_h = (\boldsymbol{v}_h, w_h)^T \ , \ \boldsymbol{v}_h \cdot \boldsymbol{n}_\Gamma = 0 \text{ on } \Gamma_n\}$. We denote I the set of vertices of the mesh, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the vertex numbered i. Then we write that $V_h = \mathrm{span}(\phi_i)_{i\in I}$ and also,

$$W_h^0 = \{k \in \{1,\dots,d+1\} = \{1,2,3\}, u_h^{(k)}(\boldsymbol{x}) = \sum_{i\in I} u_i^{(k)} \phi_i(\boldsymbol{x}) \ , \ (u_i^{(1)}, u_i^{(2)})^T \cdot \boldsymbol{n}_\Gamma = 0 \text{ for } \boldsymbol{x}_i \in \overline{\Gamma}_n\}.$$

Remark 2.4.12 (*The boundary condition on Γ_n*)

We recall that, thanks to the lifting, the boundary condition for $\boldsymbol{u}^\dagger$ on Γ_n reads $H\boldsymbol{u}^\dagger \cdot \boldsymbol{n}_s = 0$ which is equivalent to $\boldsymbol{u}^\dagger \cdot \boldsymbol{n}_s = 0$ on Γ_n thanks to hypothesis (2.5b).

At the discrete level, regarding the grid, we have to include the nodes at the end of Γ_n in order to fulfill the Neumann condition all along Γ_n. Indeed, for $\boldsymbol{x} \in \overline{\Gamma}_n$ there exists $(i,j) \in I^2$ and $\beta \in (0,1)$ such that $\boldsymbol{x} = \beta \boldsymbol{x}_i + (1-\beta)\boldsymbol{x}_j$. Then for any $\boldsymbol{u}_h = (\boldsymbol{v}_h, w_h)^T \in W_h^0$, we have,

$$\boldsymbol{v}_h(\boldsymbol{x}) \cdot \boldsymbol{n}_\Gamma(\boldsymbol{x}) = \beta \boldsymbol{v}_h(\boldsymbol{x}) \cdot \boldsymbol{n}_\Gamma(\boldsymbol{x}_i) + (1-\beta)\boldsymbol{v}_h(\boldsymbol{x}) \cdot \boldsymbol{n}_\Gamma(\boldsymbol{x}_j).$$

So it is equivalent to say $\boldsymbol{v}_h \cdot \boldsymbol{n}_\Gamma = 0$ on Γ_n as imposed $\boldsymbol{v}_h(\boldsymbol{x}_i) \cdot \boldsymbol{n}_\Gamma(\boldsymbol{x}_i) = 0$ for any $i \in I_{\Gamma_n}$, with $\boldsymbol{v}_h(\boldsymbol{x}_i) = 0$ if $\boldsymbol{x}_i$ corresponds to a boundary angle and where $I_{\Gamma_n} = \{i \in I \mid \boldsymbol{x}_i \in \overline{\Gamma}_n\}$.

In the following h denotes the *mesh size* which can be seen here as the length of the longest edge of the mesh.

AGD-Coercivity

Following the Definition 2.4.3, we are looking for a constant $C_\mathcal{D}$, independent of h such that, for any $q \in X_{\mathcal{D},\Omega,\Gamma_n}$, $\|\Pi_\mathcal{D} q\|_{L^2(\Omega)} \leq C_\mathcal{D} \|\mathrm{G}_\mathcal{D} q\|_{L^2(\Omega)^{d+1}}$.

Let $q \in X_{\mathcal{D},\Omega,\Gamma_n}$, we take $\boldsymbol{\omega}_h = (0,0,\Pi_\mathcal{D} q)^T \in W_h^0$. Thus we have, using (2.1b), $\mathrm{div}_s^\alpha(\boldsymbol{\omega}_h) = \alpha\, \Pi_\mathcal{D} q$. And by using the item (i) in the construction of $\mathrm{G}_\mathcal{D}$, we obtain,

$$\int_\Omega \mathrm{G}_\mathcal{D} q \cdot \widehat{\boldsymbol{\omega}}_h \, \mathrm{d}\boldsymbol{x} + \alpha \int_\Omega (\Pi_\mathcal{D} q)^2 \, \mathrm{d}\boldsymbol{x} = 0.$$

Then using the Cauchy-Schwarz inequality, we can write,

$$\alpha \|\Pi_\mathcal{D} q\|^2_{L^2(\Omega)} \leq \|\mathrm{G}_\mathcal{D} q\|_{L^2(\Omega)^{d+1}} \|\widehat{\boldsymbol{\omega}}_h\|_{L^2(\Omega)^{d+1}}.$$

We can remark that $\|\widehat{\boldsymbol{\omega}}_h\|_{L^2(\Omega)^{d+1}} = \|\widehat{\Pi_\mathcal{D} q}\|_{L^2(\Omega)}$ and also that,

$$\|\Pi_\mathcal{D} q\|^2_{L^2(\Omega)} = \sum_{K \in \mathcal{M}} \int_K \left(\sum_{j \in I_K} \beta_j \phi_j \right)^2 \mathrm{d}\boldsymbol{x} \quad , \quad \|\widehat{\Pi_\mathcal{D} q}\|^2_{L^2(\Omega)} = \sum_{K \in \mathcal{M}} \int_K \left(\sum_{j \in I_K} \beta_j \widehat{\phi}_j \right)^2 \mathrm{d}\boldsymbol{x} \ ,$$

where $q = (\beta_1, \ldots, \beta_N)^T$ and $I_K = \{i \in I \mid \boldsymbol{x}_i \in K\}$. But, due to the mass lumping, we can write that, for any $K \in \mathcal{M}$,

$$\int_K \left(\sum_{j \in I_K} \beta_j \widehat{\phi}_j \right)^2 \mathrm{d}\boldsymbol{x} = \int_K \left(\beta_{j_1} \widehat{\phi}_{j_1} + \beta_{j_2} \widehat{\phi}_{j_2} + \beta_{j_3} \widehat{\phi}_{j_3} \right)^2 \mathrm{d}\boldsymbol{x} = \frac{|K|}{3} (\beta_{j_1}^2 + \beta_{j_2}^2 + \beta_{j_3}^2),$$

where $I_K := \{j_1, j_2, j_3\}$ is the set of indices of the vertices of the triangular cell K. On the other hand we have, with $\boldsymbol{\beta}_K := (\beta_{j_1}, \beta_{j_2}, \beta_{j_3})^T$,

$$\int_K \left(\sum_{j \in I_K} \beta_j \phi_j \right)^2 \mathrm{d}\boldsymbol{x} \ = \ \boldsymbol{\beta}_K^T M_K \boldsymbol{\beta}_K \ ,$$

where $M_K = \dfrac{|K|}{12}\begin{pmatrix}2&1&1\\1&2&1\\1&1&2\end{pmatrix}$ is the mass matrix which is positive definite, thus satisfies,

$$\boldsymbol{\beta}_K^T M_K \boldsymbol{\beta}_K \geq \min\{\,\lambda\,,\,\lambda\in \mathrm{Spec}(M_K)\,\}\,\boldsymbol{\beta}^T\boldsymbol{\beta}.$$

So we obtain,

$$\int_K \left(\sum_{j\in I_K}\beta_j\phi_j\right)^2 \mathrm{d}\boldsymbol{x} \;\geq\; \frac{|K|}{12}(\beta_{j_1}^2+\beta_{j_2}^2+\beta_{j_3}^2).$$

Finally we have proved that,

$$\|\widehat{\Pi_{\mathcal{D}}q}\|_{L^2(\Omega)}^2 \leq 4\|\Pi_{\mathcal{D}}q\|_{L^2(\Omega)}^2.$$

Thus, we can write that,

$$\|\Pi_{\mathcal{D}}q\|_{L^2(\Omega)} \leq \frac{2}{\alpha}\|\mathrm{G}_{\mathcal{D}}q\|_{L^2(\Omega)^{d+1}},$$

which involves $C_{\mathcal{D}} = \dfrac{2}{\alpha}$.

Remark 2.4.13 (*Norm on $X_{\mathcal{D}}$*)

When the AGD-Coercivity holds, $\|\mathrm{G}_{\mathcal{D}}(\cdot)\|_{L^2(\Omega)^{d+1}}$ is a norm on $X_{\mathcal{D},\Omega,\Gamma_n}$. Indeed, let $q\in X_{\mathcal{D},\Omega,\Gamma_n}$ such that $\|\mathrm{G}_{\mathcal{D}}q\|_{L^2(\Omega)^{d+1}}=0$, thus $\|\Pi_{\mathcal{D}}q\|_{L^2(\Omega)}=0$. Since the family $(\phi_i)_{i\in I}$ is linearly independent, we infer that $q=0$.

AGD-Limit-conformity

As mentioned in [22, Lemma A.9], it is sufficient to prove the limit-conformity in a dense subset of $H^0_{\mathrm{div}_s^\alpha}(\Omega)$. Thus we take, $\boldsymbol{\omega}\in H^0_{\mathrm{div}_s^\alpha}(\Omega)\cap(\mathcal{C}^2(\overline{\Omega}))^{d+1}$. Then we denote $\boldsymbol{\omega} := (\omega^{(1)},\omega^{(2)},\omega^{(3)})^T$ and $\boldsymbol{\omega}_h = (\omega_h^{(1)},\omega_h^{(2)},\omega_h^{(3)})^T$ where, for $k\in\{1,2,3\}$,

$$\omega_h^{(k)} \;:=\; \sum_{i\in I}\omega^{(k)}(\boldsymbol{x}_i)\;\phi_i.$$

We have $\boldsymbol{\omega}_h\in W_h^0$ since $\boldsymbol{\omega}_h\cdot\boldsymbol{n}_s=0$ (due to the fact that $\boldsymbol{\omega}\cdot\boldsymbol{n}_s=0$).

Then we can use the item (i) in the construction of $\mathrm{G}_{\mathcal{D}}$ to write,

$$\int_\Omega \mathrm{G}_{\mathcal{D}}q\cdot\boldsymbol{\omega}+\Pi_{\mathcal{D}}q\,\mathrm{div}_s^\alpha\boldsymbol{\omega}\;\mathrm{d}\boldsymbol{x} = \int_\Omega \mathrm{G}_{\mathcal{D}}q\cdot(\boldsymbol{\omega}-\widehat{\boldsymbol{\omega}}_h)+\Pi_{\mathcal{D}}q\,\mathrm{div}_s^\alpha(\boldsymbol{\omega}-\boldsymbol{\omega}_h)\;\mathrm{d}\boldsymbol{x}\;,\;\forall q\in X_{\mathcal{D},\Omega,\Gamma_n}.$$

Using the Cauchy-Schwarz inequality and the AGD-Coercivity, we obtain,

$$\left|\int_\Omega \mathrm{G}_{\mathcal{D}}q\cdot\boldsymbol{\omega}+\Pi_{\mathcal{D}}q\,\mathrm{div}_s^\alpha\boldsymbol{\omega}\;\mathrm{d}\boldsymbol{x}\right| \leq \left(\,\|\widehat{\boldsymbol{\omega}}_h-\boldsymbol{\omega}_h\|_{L^2(\Omega)^{d+1}}+\|\boldsymbol{\omega}-\boldsymbol{\omega}_h\|_{H^0_{\mathrm{div}_s^\alpha}(\Omega)}(1+C_{\mathcal{D}})\,\right)\|q\|_{\mathcal{D}}\;,\;\forall q\in X_{\mathcal{D},\Omega,\Gamma_n}.$$

Component by component, for $k\in\{1,2,3\}$, we get (see for instance equation (8.18) in [22]) that,

$$\|\widehat{\boldsymbol{\omega}}_h^{(k)}-\boldsymbol{\omega}_h^{(k)}\|_{L^2(\Omega)} \leq h\,\|\nabla\boldsymbol{\omega}_h^{(k)}\|_{L^2(\Omega)^d}\;,$$

where $\nabla\boldsymbol{\omega}_h$ remains bounded since $\|\nabla\widehat{\boldsymbol{\omega}}^{(k)}-\nabla\boldsymbol{\omega}_h^{(k)}\|_{L^2(\Omega)^d}\leq h\|\boldsymbol{\omega}^{(k)}\|_{2,\infty}$, with $\|\boldsymbol{\omega}^{(k)}\|_{2,\infty}=\sup_\Omega|\partial^2\boldsymbol{\omega}^{(k)}|$. We can also write,

$$\|\boldsymbol{\omega}^{(k)}-\boldsymbol{\omega}_h^{(k)}\|_{H^1(\Omega)} \leq h\|\boldsymbol{\omega}^{(k)}\|_{2,\infty}\;,\;\forall k\in\{1,2,3\}.$$

Moreover, by definition of the norm on $H^0_{\mathrm{div}^\alpha_s}(\Omega)$ given by (2.40), we have $C > 0$ independent of the mesh such that,

$$\|\boldsymbol{\omega} - \boldsymbol{\omega}_h\|_{H^0_{\mathrm{div}^\alpha_s}(\Omega)} \leq C \, \|\boldsymbol{\omega} - \boldsymbol{\omega}_h\|_{H^1(\Omega)^d},$$

which allows us to conclude that our AGD is limit-conforming in the sens of Definition 2.4.5.

AGD-Consistency

Following the Definition 2.4.4, we have to prove that, $\forall \psi \in H^1_d(\Omega)$,

$$\lim_{h\to 0} S_{\mathcal{D}_h}(\psi) = \lim_{h\to 0} \min_{q\in X_{\mathcal{D},\Omega,\Gamma_n}} \Big(\|\Pi_{\mathcal{D}h} q - \psi\|_{L^2(\Omega)} + \|\mathrm{G}_{\mathcal{D}_h} q - \nabla^\alpha_s \psi\|_{L^2(\Omega)^{d+1}} \Big) = 0.$$

As mentioned in [22, Lemma A.13], it is sufficient to prove this result in a dense subset of $H^1_d(\Omega)$. Thus we will take, $\psi \in H^1_d(\Omega) \cap \mathcal{C}^2(\overline{\Omega})$.

We have, for all $q \in X_{\mathcal{D},\Omega,\Gamma_n}$,

$$0 \leq S_{\mathcal{D}}(\psi) \leq \|\Pi_{\mathcal{D}h} q - \psi\|_{L^2(\Omega)} + \|\mathrm{G}_{\mathcal{D}_h} q - \nabla^\alpha_s \psi\|_{L^2(\Omega)^{d+1}}.$$

We introduce $q = (\psi(\boldsymbol{x}_i))_{i\in I}$, thus $q \in X_{\mathcal{D},\Omega,\Gamma_n}$ and $\Pi_{\mathcal{D}} q = \sum_{i\in I} \psi(\boldsymbol{x}_i)\phi_i$.

By definition of Q_h, $\Pi_{\mathcal{D}} q$ is the $\mathbb{P}_1$ finite element approximation of ψ, thus we have,

$$\lim_{h\to 0} \|\Pi_{\mathcal{D}} q - \psi\|_{L^2(\Omega)} = 0.$$

Now we have to study the term $\|\mathrm{G}_{\mathcal{D}} q - \nabla^\alpha_s \psi\|_{L^2(\Omega)^{d+1}}$.

Thanks to the item (ii) in the construction of $\mathrm{G}_{\mathcal{D}}$, we know that there exists $G_h \in W^0_h$ such that $\mathrm{G}_{\mathcal{D}} q = H\widehat{G}_h$. For any $i \in I$, and almost everywhere in K_i (the dual cell around the vertex i), the function $\widehat{G}_h$ is constant and, in the following, we denote this quantity $\widehat{G}_{h,i}$.

We are looking for a bound of the difference $\mathrm{G}_{\mathcal{D}} q - \nabla^\alpha_s \psi$. For all $\boldsymbol{x} \in \overline{\Omega}$, there exists $i \in I$ such that $\boldsymbol{x} \in K_i$ and we can write,

$$\mathrm{G}_{\mathcal{D}} q(\boldsymbol{x}) - \nabla^\alpha_s \psi(\boldsymbol{x}) = (H(\boldsymbol{x}_i)\,\widehat{G}_{h,i} - \nabla^\alpha_s \psi(\boldsymbol{x}_i)) + (H(\boldsymbol{x}) - H(\boldsymbol{x}_i))\,\widehat{G}_{h,i} - (\nabla^\alpha_s \psi(\boldsymbol{x}) - \nabla^\alpha_s \psi(\boldsymbol{x}_i)).$$

Since for all $(a,b,c) \in \mathbb{R}^3$, we have $(a+b+c)^2 \leq 3\,(a^2+b^2+c^2)$, we can write,

$$\begin{aligned} \|\mathrm{G}_{\mathcal{D}} q - \nabla^\alpha_s \psi\|^2_{L^2(\Omega)^{d+1}} \leq 3 \sum_{i\in I} \Big(& \|H(\boldsymbol{x}_i)\widehat{G}_{h,i} - \nabla^\alpha_s \psi(\boldsymbol{x}_i)\|^2_{L^2(K_i)^{d+1}} + \\ & \|(H - H(\boldsymbol{x}_i))\,\widehat{G}_{h,i}\|^2_{L^2(K_i)^{d+1}} + \|\nabla^\alpha_s \psi - \nabla^\alpha_s \psi(\boldsymbol{x}_i)\|^2_{L^2(K_i)^{d+1}} \Big). \end{aligned} \quad (2.55)$$

Now we have to study the three terms inside the sum. To do so we have to compute $\widehat{G}^{(k)}_{h,i}$ for any vertex i and any coordinate $k \in \{1,2,3\}$. To do so, we need to distinguish the value of k, but also if the vertex i is, or not, on Γ_n. This is why we distinguish if $\boldsymbol{x}_i$ belongs, or not, to Γ_n.

• If $\boldsymbol{x}_i \in \overline{\Omega} \setminus \overline{\Gamma}_n$.

[I] We take $\boldsymbol{\omega}_h = (\phi_i, 0, 0)^T$. By localisation of the vertex i, we have $\boldsymbol{\omega}_h = 0$ on Γ_n. Thus $\boldsymbol{\omega}_h \in W_h^0$. Moreover $\int_\Omega \mathrm{G}_\mathcal{D} q \cdot \widehat{\boldsymbol{\omega}}_h \, \mathrm{d}\boldsymbol{x} = \int_{K_i} \mathrm{G}_\mathcal{D}^{(1)} q \, \mathrm{d}\boldsymbol{x}$ where K_i still denotes the dual cell around the vertex i. We can write,

$$\int_\Omega \mathrm{G}_\mathcal{D} q \cdot \widehat{\boldsymbol{\omega}}_h \, \mathrm{d}\boldsymbol{x} = H_i \, |K_i| \, \widehat{G}_{h,i}^{(1)} \ ,$$

with $H_i = \frac{1}{|K_i|} \int_{K_i} H \, \mathrm{d}\boldsymbol{x}$. Then, using the item (i) in the construction of $\mathrm{G}_\mathcal{D}$ and the duality relation (2.27) (since $\Pi_\mathcal{D} q \in H_d^1(\Omega)$ and $\boldsymbol{\omega}_h \in W_h^0 \subset H_{\mathrm{div}_s^\alpha}(\Omega)$), we can write,

$$H_i \, |K_i| \, \widehat{G}_{h,i}^{(1)} = \int_\Omega \boldsymbol{\omega}_h \cdot \nabla_s^\alpha \Pi_\mathcal{D} q \, \mathrm{d}\boldsymbol{x} \ ,$$

which implies by definitions of $\boldsymbol{\omega}_h$ and ∇_s^α (2.1a),

$$H_i |K_i| \widehat{G}_{h,i}^{(1)} = \sum_{T \in \mathcal{M}_i} \Big(\int_T \phi_i \, (\, H \, (\nabla \Pi_\mathcal{D} q)^{(1)} + \Pi_\mathcal{D} q \, (\nabla \zeta)^{(1)} \,) \, \mathrm{d}\boldsymbol{x} \, \Big) \ ,$$

where $\mathcal{M}_i$ is the set of triangles of the mesh connected to the vertex i. We can find in [27] that, for every $T \in \mathcal{M}_i$,

$$\| (\nabla \Pi_\mathcal{D} q)^{(1)} - (\nabla \psi)^{(1)} \|_{L^\infty(T)} \leq h \, \|\psi\|_{2,\infty} \ ,$$

and

$$\| \Pi_\mathcal{D} q - \psi \|_{L^\infty(T)} \leq h \, \|\psi\|_{2,\infty} \ .$$

Thus, we can introduce R_1 and R_2, two bounded functions defined on $\bigcup_{T \in \mathcal{M}_i} T$ such that,

$$(\nabla \Pi_\mathcal{D} q)^{(1)}(\boldsymbol{x}) = (\nabla \psi)^{(1)}(\boldsymbol{x}_i) + h \, R_1(\boldsymbol{x}),$$

and

$$\Pi_\mathcal{D} q(\boldsymbol{x}) = \psi(\boldsymbol{x}_i) + h \, R_2(\boldsymbol{x}),$$

thus with $|R_l(\boldsymbol{x})| \leq \|\psi\|_{2,\infty}$ for $l = 1, 2$ and $\boldsymbol{x} \in \bigcup_{T \in \mathcal{M}_i} T$. This implies,

$$\sum_{T \in \mathcal{M}_i} \int_T \phi_i \left(H (\nabla \Pi_\mathcal{D} q)^{(1)} + \Pi_\mathcal{D} q (\nabla \zeta)^{(1)} \right) \mathrm{d}\boldsymbol{x} = \sum_{T \in \mathcal{M}_i} \left[\int_T \phi_i \, (\, H \, (\nabla \psi)^{(1)}(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i) (\nabla \zeta)^{(1)} \,) \, \mathrm{d}\boldsymbol{x} + h \, \mathcal{J}_T \right].$$

where $\mathcal{J}_T = \int_T \phi_i H R_1 + (\nabla \zeta)^{(1)} R_2 \, \mathrm{d}\boldsymbol{x}$. Since $0 \leq \phi_i \leq 1$ and the assumptions on H (2.5b) and ζ (2.5c), all the terms inside the integral $\mathcal{J}_T$ are bounded on T and thus $|\mathcal{J}_T| \leq |T| C$ with C independent of the mesh.

So we can introduce R_3 a bounded function defined on $\bigcup_{T \in \mathcal{M}_i} T$ such that,

$$H_i \, |K_i| \, \widehat{G}_{h,i}^{(1)} = \sum_{T \in \mathcal{M}_i} \left[\int_T \phi_i \, \left(\, H \, (\nabla \psi)^{(1)}(\boldsymbol{x}_i) + \psi(x_i) (\nabla \zeta)^{(1)} \, \right) \mathrm{d}\boldsymbol{x} \right] + h \, |K_i| \, R_3 \ .$$

Then we write that, for $T \in \mathcal{M}_i$ and $\boldsymbol{x} \in T$,

$$\begin{aligned} H(\boldsymbol{x}) (\nabla \psi)^{(1)}(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i) (\nabla \zeta)^{(1)}(\boldsymbol{x}) &= H(\boldsymbol{x}_i) (\nabla \psi)^{(1)}(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i) (\nabla \zeta)^{(1)}(\boldsymbol{x}_i) \\ &\quad + (H(\boldsymbol{x}) - H(\boldsymbol{x}_i)) (\nabla \psi)^{(1)}(\boldsymbol{x}_i) + ((\nabla \zeta)^{(1)}(\boldsymbol{x}) - (\nabla \zeta)^{(1)}(\boldsymbol{x}_i)) \psi(\boldsymbol{x}_i) \ . \end{aligned} \quad (2.56)$$

Due to the assumptions on H (2.5b) and ζ (2.5c), we write that there exists a bounded function R_4 defined on $\bigcup_{T\in\mathcal{M}_i} T$ such that,

$$H(\boldsymbol{x})(\nabla\psi)^{(1)}(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i)(\nabla\zeta)^{(1)}(\boldsymbol{x}) = H(\boldsymbol{x}_i)(\nabla\psi)^{(1)}(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i)(\nabla\zeta)^{(1)}(\boldsymbol{x}_i) + h\,R_4(\boldsymbol{x}).$$

Since $\sum_{T\in\mathcal{M}_i}\int_T \phi_i(x)\,\mathrm{d}\boldsymbol{x} = |K_i|$, we obtain,

$$H_i\,|K_i|\,\widehat{G}^{(1)}_{h,i} = |K_i|(\underbrace{H(\boldsymbol{x}_i)(\nabla\psi)^{(1)}(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i)(\nabla\zeta)^{(1)}(\boldsymbol{x}_i)}_{=(\nabla^\alpha_s\psi)^{(1)}(\boldsymbol{x}_i)} + h\,(R_3+R_4)\,).$$

Since $|H_i - H(\boldsymbol{x}_i)| \le C\,h$ for a constant $C>0$, we can prove that,

$$|H(\boldsymbol{x}_i)\widehat{G}^{(1)}_{h,i} - \nabla^\alpha_s\psi(\boldsymbol{x}_i)^{(1)}| \le h\,C_*\ ,$$

and thus, there exists $R_*>0$ independent of the mesh such that,

$$\|(\mathrm{G}_{\mathcal{D}}q)^{(1)} - (\nabla^\alpha_s\psi)^{(1)}\|_{L^\infty(\Omega_{int})} \le h\,R_*,$$

where $\Omega_{int} = \bigcup_{T\in\mathcal{M}_i|i\in I_{int}} T$ with $I_{int} = \{i\in I \mid \boldsymbol{x}_i \notin \overline{\Gamma}_n\}$. We can write,

$$\|(\mathrm{G}_{\mathcal{D}}q)^{(1)} - (\nabla^\alpha_s\psi)^{(1)}\|^2_{L^2(\Omega_{int})} \le |\Omega_{int}|\,C\,h^2. \tag{2.57}$$

[II] If we take $\boldsymbol{\omega}_h = (0,\phi_i,0)^T$ a similar demonstration proves that,

$$\|(\mathrm{G}_{\mathcal{D}}q)^{(2)} - (\nabla^\alpha_s\psi)^{(2)}\|_{L^\infty(\Omega_{int})} \le h\,R_*,$$

and thus,

$$\|(\mathrm{G}_{\mathcal{D}}q)^{(2)} - (\nabla^\alpha_s\psi)^{(2)}\|^2_{L^2(\Omega_{int})} \le |\Omega_{int}|\,C\,h^2 \tag{2.58}$$

[III] If we take $\boldsymbol{\omega}_h = (0,0,\phi_i)^T$, by following the same strategy, we obtain,

$$\begin{aligned} H_i\,|K_i|\,(G_h)^{(3)}_i &= \int_\Omega \boldsymbol{\omega}_h\cdot\nabla^\alpha_s\Pi_{\mathcal{D}}q\,\mathrm{d}\boldsymbol{x} = \int_\Omega \phi_i(\nabla^\alpha_s\Pi_{\mathcal{D}}q)^{(3)}\,\mathrm{d}\boldsymbol{x} = \int_\Omega \phi_i(-\alpha\Pi_{\mathcal{D}}q)\,\mathrm{d}\boldsymbol{x}\\ &= -\alpha\int_\Omega \phi_i\Big(\sum_{j\in I}\psi(\boldsymbol{x}_j)\phi_j\Big)\,\mathrm{d}\boldsymbol{x} = -\alpha\sum_{T\in\mathcal{M}_i}\Big(\int_T \phi_i\sum_{j\in I_T}\psi(\boldsymbol{x}_j)\phi_j\,\mathrm{d}\boldsymbol{x}\Big)\\ &= -\alpha\sum_{T\in\mathcal{M}_i}\int_T \phi_i\Big(\sum_{j\in I_T}\psi(\boldsymbol{x}_i)\phi_j + \sum_{j\in I_T}(\psi(\boldsymbol{x}_j)-\psi(\boldsymbol{x}_i))\phi_j\Big)\,\mathrm{d}\boldsymbol{x}, \end{aligned}$$

where $I_T = \{i\in I \mid \boldsymbol{x}_i\in T\}$. We have,

$$\sum_{j\in I_T}\psi(\boldsymbol{x}_i)\phi_j = \psi(\boldsymbol{x}_i)\underbrace{\sum_{j\in I_T}\phi_j}_{=1} = \psi(\boldsymbol{x}_i).$$

We have also $0\le\phi_j\le 1$ and, due to the regularity of ψ there exists R_T, a bounded function defined on T, such that $(\psi(\boldsymbol{x}_j)-\psi(\boldsymbol{x}_i))\phi_j = h\,R_T$. Since $(\nabla^\alpha_s\psi)^{(3)}(\boldsymbol{x}_i) = -\alpha\psi(\boldsymbol{x}_i)$ and $\sum_{T\in\mathcal{M}_i}\int_T\phi_i\,\mathrm{d}\boldsymbol{x} = |K_i|$, we can write,

$$|H_i(G_h)^{(3)}_i - (\nabla^\alpha_s\psi)^{(3)}| \le C\,h.$$

And, by using the same trick between H_i and $H(\boldsymbol{x}_i)$, we can write,

$$|H(\boldsymbol{x}_i)\widehat{G}_{h,i}^{(3)} - (\nabla_s^\alpha \psi(\boldsymbol{x}_i))^{(3)}| \leq h\, C_* \ ,$$

and thus,

$$\|(\mathrm{G}_{\mathcal{D}} q)^{(3)} - (\nabla_s^\alpha \psi)^{(3)}\|_{L^\infty(\Omega_{int})} \leq h\, C_\star.$$

We remark that these computations are independent of the assumption $\boldsymbol{x}_i \in \overline{\Omega} \setminus \overline{\Gamma}_n$. Thus we can obtain the same estimation if $\boldsymbol{x}_i \in \overline{\Gamma}_n$ and we can write,

$$\|(\mathrm{G}_{\mathcal{D}} q)^{(3)} - (\nabla_s^\alpha \psi)^{(3)}\|_{L^2(\Omega)}^2 \leq |\Omega|\, C\, h^2. \tag{2.59}$$

• If $\boldsymbol{x}_i \in \overline{\Gamma}_n$.

We take $\boldsymbol{\omega}_h = (\beta\phi_i, \lambda\phi_i, 0)^T$ (implicitly β and λ depend of the vertex i).
Thus $\boldsymbol{\omega}_h(\boldsymbol{x}) \cdot \boldsymbol{n}_{\Gamma_n}(\boldsymbol{x}) = \beta \boldsymbol{n}_{\Gamma_n}^{(1)}(\boldsymbol{x}) + \lambda \boldsymbol{n}_{\Gamma_n}^{(2)}(\boldsymbol{x})$ for $\boldsymbol{x} \in \overline{\Gamma}_n$. So in order to obtain $\boldsymbol{\omega}_h \in W_h^0$, we need to have (see Remark 2.4.12) $\beta \boldsymbol{n}_{\Gamma_n}^{(1)}(\boldsymbol{x}_i) + \lambda \boldsymbol{n}_{\Gamma_n}^{(2)}(\boldsymbol{x}_i) = 0$. For instance, we can chose $\beta = \boldsymbol{n}_{\Gamma_n}^{(2)}(\boldsymbol{x}_i) := \boldsymbol{n}_{\Gamma_n,i}^{(2)}$ and $\lambda = -\boldsymbol{n}_{\Gamma_n}^{(1)}(\boldsymbol{x}_i) := -\boldsymbol{n}_{\Gamma_n,i}^{(1)}$.

As recalled, we have $\mathrm{G}_{\mathcal{D}} q = H\widehat{G}_h$ with $G_h \in W_h^0$, thus $G_h \cdot \boldsymbol{n}_{\Gamma_n} = 0$. Since $(\widehat{G}_h)_i = G_h(\boldsymbol{x}_i)$ we can write $(\widehat{G}_h)_i^{(1)} \boldsymbol{n}_{\Gamma_n,i}^{(1)} + (\widehat{G}_h)_i^{(2)} \boldsymbol{n}_{\Gamma_n,i}^{(2)} = 0$.

Moreover,

$$\int_\Omega \mathrm{G}_{\mathcal{D}} q \cdot \widehat{\boldsymbol{\omega}}_h \ \mathrm{d}\boldsymbol{x} = \int_{K_i} \mathrm{G}_{\mathcal{D}} q \cdot \begin{pmatrix} \beta \\ \lambda \\ 0 \end{pmatrix} \ \mathrm{d}\boldsymbol{x},$$

where K_i still denotes the dual cell around the vertex i. We can write,

$$\int_\Omega \mathrm{G}_{\mathcal{D}} q \cdot \widehat{\boldsymbol{\omega}}_h \ \mathrm{d}\boldsymbol{x} = H_i\, |K_i|\, \widehat{G}_{h,i} \cdot \begin{pmatrix} \beta \\ \lambda \\ 0 \end{pmatrix} \ \mathrm{d}\boldsymbol{x},$$

with $H_i = \frac{1}{|K_i|} \int_{K_i} H \ \mathrm{d}\boldsymbol{x}$. Then using the item (i) in the construction of $\mathrm{G}_{\mathcal{D}}$ and the duality relation (2.27) (since $\Pi_{\mathcal{D}} q \in H_d^1(\Omega)$ and $\boldsymbol{\omega}_h \in W_h^0 \subset H_{\mathrm{div}_s^\alpha}(\Omega)$), we can write,

$$H_i\, |K_i|\, \widehat{G}_{h,i} \cdot \begin{pmatrix} \beta \\ \lambda \\ 0 \end{pmatrix} = \int_\Omega \boldsymbol{\omega}_h \cdot \nabla_s^\alpha \Pi_{\mathcal{D}} q \ \mathrm{d}\boldsymbol{x},$$

which implies, by definitions of $\boldsymbol{\omega}_h$ and ∇_s^α (2.1a),

$$H_i\, |K_i|\, \widehat{G}_{h,i} \cdot \begin{pmatrix} \beta \\ \lambda \\ 0 \end{pmatrix} = \sum_{T \in \mathcal{M}_i} \int_T \phi_i \ \big(H(\nabla \Pi_{\mathcal{D}} q) + \Pi_{\mathcal{D}} q (\nabla \zeta) \big) \cdot \begin{pmatrix} \beta \\ \lambda \end{pmatrix} \ \mathrm{d}\boldsymbol{x},$$

where $\mathcal{M}_i$ is the set of triangles of the mesh connected to the vertex i. We can reused that, for $k = 1, 2$,

$$\|(\nabla \Pi_{\mathcal{D}} q)^{(k)} - (\nabla \psi)^{(k)}\|_{L^\infty(T)} \leq h \|\psi\|_{2,\infty},$$

and

$$\|\Pi_{\mathcal{D}} q - \psi\|_{L^\infty(T)} \leq h \|\psi\|_{2,\infty}.$$

Thus, we can introduce $\boldsymbol{R}_1$ and R_2 two bounded functions defined on $\bigcup_{T\in\mathcal{M}_i} T$ such that,

$$(\nabla \Pi_{\mathcal{D}} q)(\boldsymbol{x}) = (\nabla\psi)(\boldsymbol{x}_i) + h\, \boldsymbol{R}_1(\boldsymbol{x}),$$

and

$$\Pi_{\mathcal{D}} q(\boldsymbol{x}) = \psi(\boldsymbol{x}_i) + h\, R_2(\boldsymbol{x}),$$

thus with $\max_{l=1,2} \|\boldsymbol{R}_1^{(l)}\|_{L^\infty(\bigcup_{T\in\mathcal{M}_i} T)} \leq \|\psi\|_{2,\infty}$ and $|R_2(\boldsymbol{x})| \leq \|\psi\|_{2,\infty}$ for $\boldsymbol{x} \in \bigcup_{T\in\mathcal{M}_i} T$. This involves,

$$\sum_{T\in\mathcal{M}_i} \int_T \phi_i \ \left(H(\nabla \Pi_{\mathcal{D}} q) + \Pi_{\mathcal{D}} q (\nabla\zeta) \right) \cdot \begin{pmatrix} \beta \\ \lambda \end{pmatrix} \ \mathrm{d}\boldsymbol{x} = \sum_{T\in\mathcal{M}_i} \left[\int_T \phi_i \ \left(H\,(\nabla\psi)(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i)(\nabla\zeta) \right) \cdot \begin{pmatrix} \beta \\ \lambda \end{pmatrix} \ \mathrm{d}\boldsymbol{x} + h\, \mathcal{J}_T \right],$$

with,

$$\mathcal{J}_T = \int_T \phi_i \ \left(H \boldsymbol{R}_1 + R_2(\nabla\zeta) \right) \cdot \begin{pmatrix} \beta \\ \lambda \end{pmatrix} \ \mathrm{d}\boldsymbol{x}.$$

Since $0 \leq \phi_i \leq 1$ and the assumptions on H (2.5b) and ζ (2.5c), all the terms inside the integral $\mathcal{J}_T$ are bounded on T and thus $\|\mathcal{J}_T\|_{L^\infty(T)} \leq |T|\, C$ with C independent of the mesh.

So we can introduce R_3 a bounded function defined on $\bigcup_{T\in\mathcal{M}_i} T$ such that,

$$H_i\, |K_i|\, \widehat{G}_{h,i} \cdot \begin{pmatrix} \beta \\ \lambda \\ 0 \end{pmatrix} = \sum_{T\in\mathcal{M}_i} \left[\int_T \phi_i \ \left(H\,(\nabla\psi)(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i)(\nabla\zeta) \right) \cdot \begin{pmatrix} \beta \\ \lambda \end{pmatrix} \ \mathrm{d}\boldsymbol{x} \right] + h\, |K_i|\, R_3.$$

Then we write that, for $T \in \mathcal{M}_i$ and $\boldsymbol{x} \in T$,

$$\begin{aligned} \left(H(\boldsymbol{x})(\nabla\psi)(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i)(\nabla\zeta)(\boldsymbol{x}) \right) \cdot \begin{pmatrix} \beta \\ \lambda \end{pmatrix} = & \\ \left(H(\boldsymbol{x}_i)(\nabla\psi)(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i)(\nabla\zeta)(\boldsymbol{x}_i) \right) \cdot \begin{pmatrix} \beta \\ \lambda \end{pmatrix} + (H(\boldsymbol{x}) - H(\boldsymbol{x}_i))(\nabla\psi)(\boldsymbol{x}_i) \cdot \begin{pmatrix} \beta \\ \lambda \end{pmatrix} & \\ + \psi(\boldsymbol{x}_i)((\nabla\zeta)(\boldsymbol{x}) - (\nabla\zeta)(\boldsymbol{x}_i)) \cdot \begin{pmatrix} \beta \\ \lambda \end{pmatrix}. & \quad (2.60) \end{aligned}$$

Due to the assumptions on H (2.5b) and ζ (2.5c), there exists a bounded function R_4 defined on $\bigcup_{T\in\mathcal{M}_i} T$ such that,

$$\left(H(\boldsymbol{x})(\nabla\psi)(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i)(\nabla\zeta)(\boldsymbol{x}) \right) \cdot \begin{pmatrix} \beta \\ \lambda \end{pmatrix} = \left(H(\boldsymbol{x}_i)(\nabla\psi)(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i)(\nabla\zeta)(\boldsymbol{x}_i) \right) \cdot \begin{pmatrix} \beta \\ \lambda \end{pmatrix} + hR_4(\boldsymbol{x}).$$

Since $\sum_{T\in\mathcal{M}_i} \int_T \phi_i(x)\ \mathrm{d}\boldsymbol{x} = |K_i|$, we obtain,

$$H_i\, |K_i|\, \widehat{G}_{h,i} \cdot \begin{pmatrix} \beta \\ \lambda \\ 0 \end{pmatrix} = |K_i| [\ \underbrace{(H(\boldsymbol{x}_i)(\nabla\psi)(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i)(\nabla\zeta)(\boldsymbol{x}_i))}_{= ((\nabla_s^\alpha\psi)^{(1)}(\boldsymbol{x}_i), (\nabla_s^\alpha\psi)^{(2)}(\boldsymbol{x}_i))^T} \cdot \begin{pmatrix} \beta \\ \lambda \end{pmatrix} \ + \ h(R_3 + R_4)\ ].$$

We can introduce $R_{\star,i} > 0$ such that,

$$\left(H_i \widehat{G}_{h,i} - \nabla_s^\alpha \psi(\boldsymbol{x}_i)\right) \cdot \begin{pmatrix} \beta \\ \lambda \\ 0 \end{pmatrix} = h\, R_{\star,i}.$$

As mentioned previously, we can chose $\beta = \boldsymbol{n}^{(2)}_{\Gamma_n}(\boldsymbol{x}_i) := \boldsymbol{n}^{(2)}_{\Gamma_n,i}$ and $\lambda = -\boldsymbol{n}^{(1)}_{\Gamma_n}(\boldsymbol{x}_i) := -\boldsymbol{n}^{(1)}_{\Gamma_n,i}$ with $\beta^2 + \lambda^2 = 1$ since $\boldsymbol{n}_{\Gamma_n}$ is a unit vector. So we have the system,

$$\left\{ \begin{array}{rcl} \beta^2 + \lambda^2 & = & 1\ , \\ \beta \widehat{G}^{(1)}_{h,i} + \lambda \widehat{G}^{(2)}_{h,i} & = & \frac{1}{H_i}\left(\beta (\nabla_s^\alpha \psi)^{(1)}(x_i) + \lambda (\nabla_s^\alpha \psi)^{(2)}(x_i) + h R_{\star,i}\right)\ , \\ -\lambda \widehat{G}^{(1)}_{h,i} + \beta \widehat{G}^{(2)}_{h,i} & = & 0\ . \end{array} \right.$$

Due to the assumptions on H (2.5b) and ζ (2.5c), the regularity of ψ, we can deduce that there exists C_0 independent of the mesh such that, for $k = 1, 2$,

$$|\widehat{G}^{(k)}_{h,i}| \leq C_0. \tag{2.61}$$

We close the discussion on the case $x_i \in \overline{\Gamma}_n$ and come back to establish an upper bound of (2.55).

- Due to the regularity of H, we have, for $\boldsymbol{x} \in K_i$, $|H(\boldsymbol{x}) - H(\boldsymbol{x}_i)| \leq h\, C$ with $C > 0$ independent of the mesh, and thus $\forall i \in I$,

  $$\|(H - H(\boldsymbol{x}_i))\widehat{G}_{h,i}\|_{L^2(K_i)^{d+1}} \leq \sqrt{|K_i|}\, \max_i \|\widehat{G}_{h,i}\|_2\, C\, h.$$

  The previous computations allow us to say that $|\widehat{G}^{(k)}_{h,i}|$ is bounded for every node $i \in I$ and for $k \in \{1, 2, 3\}$. So we can write,

  $$\sum_{i \in I} \|(H - H(\boldsymbol{x}_i))\widehat{G}_{h,i}\|^2_{L^2(K_i)^{d+1}} \leq |\Omega|\ C_1\ h^2.$$

- Now we use the definition of ∇_s^α (2.1a), and we have, for all $i \in I$ and $\boldsymbol{x} \in K_i$,

  $$\nabla_s^\alpha \psi(\boldsymbol{x}) - \nabla_s^\alpha \psi(\boldsymbol{x}_i) = (H(\boldsymbol{x})\nabla\psi(\boldsymbol{x}) - H(\boldsymbol{x}_i)\nabla\psi(\boldsymbol{x}_i) + \psi(\boldsymbol{x})\nabla\zeta(\boldsymbol{x}) - \psi(\boldsymbol{x}_i)\nabla\zeta(\boldsymbol{x}_i), -\alpha(\psi(\boldsymbol{x}) - \psi(\boldsymbol{x}_i))^T.$$

  Due to the regularity of H, ψ, ζ, we can write,

  $$\sum_{i \in I} \|\nabla_s^\alpha \psi - \nabla_s^\alpha \psi(\boldsymbol{x}_i)\|^2_{L^2(K_i)^{d+1}} \leq |\Omega|\ C_2\ h^2.$$

- Finally, we have for any $i \in I$,

  $$\|(H(\boldsymbol{x}_i)\widehat{G}_{h,i} - \nabla_s^\alpha \psi(\boldsymbol{x}_i))\|^2_{L^2(K_i)^{d+1}} = \sum_{k=1,2,3} \|(H(\boldsymbol{x}_i)\widehat{G}^{(k)}_{h,i} - (\nabla_s^\alpha \psi)^{(k)}(\boldsymbol{x}_i))\|^2_{L^2(K_i)}.$$

  By (2.59), we can write,

  $$\sum_{i \in I} \|(H(\boldsymbol{x}_i)\widehat{G}^{(3)}_{h,i} - (\nabla_s^\alpha \psi)^{(3)}(\boldsymbol{x}_i))\|^2_{L^2(K_i)} \leq |\Omega|\ C_3\ h^2.$$

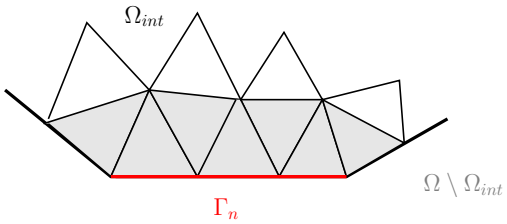
And, thanks to (2.57) and (2.58), we can write,

$$\sum_{i\in I \ | \ \boldsymbol{x}_i\notin\overline{\Gamma}_n} \sum_{k=1,2} \|(H(\boldsymbol{x}_i)\widehat{G}_{h,i}^{(k)} - (\nabla_s^\alpha\psi)^{(k)}(\boldsymbol{x}_i))\|_{L^2(K_i)}^2 \leq |\Omega_{int}|\ C_4\ h^2.$$

And by using (2.61), and the assumptions on data, we can write,

$$\sum_{i\in I \ | \ \boldsymbol{x}_i\in\overline{\Gamma}_n} \sum_{k=1,2} \|(H(\boldsymbol{x}_i)\widehat{G}_{h,i}^{(k)} - (\nabla_s^\alpha\psi)^{(k)}(\boldsymbol{x}_i))\|_{L^2(K_i)}^2 \leq \underbrace{|\Omega\setminus\Omega_{int}|}_{\leq(\ |\Gamma_n|\ +\ m\ )\ h}\ C_4,$$

where $m \geq 0$ is associated to the geometry of Γ_n. The figure below can help to understand the bound of $|\Omega \setminus \Omega_{int}|$.



2.5 Conclusion

In this chapter, some novel results and open questions have been exhibited regarding the analysis of the elliptic problem (2.2) and the GDM framework.

- In Section 2.2 a new conforming formulation the only unknown is the pressure has been exhibited, the velocity being post-processed. An error estimate has been proved thanks to the GDM framework. The $\mathbb{P}_1$ finite element method is encompassed in this framework. A future work could be to implement this strategy in order to compare it with the mixed strategy. The comparison could be done in terms of accuracy and numerical cost by paying attention to the cost of the linear solver and, if applicable, to the chosen preconditioner. The application of this novel strategy to simulate the full model (including the hyperbolic part) implies to solve the problem of handling the boundary conditions in the full model and the localisation of the unknowns (some reconstructions have to be done due to the splitting technique).

- In Section 2.3, as far as we know, it is the first time that the Abstract GDM framework is applied to analyse *special* operators. Moreover we have seen that this framework allows us (contrary to section 2.2) to obtain error estimates even if the function reconstruction used in the *special* gradient is not the same as in the GD-Definition.

- In Section 2.4 we have seen how the two schemes used in [2] can be interpreted as Gradient Schemes. We focused, at the end of this chapter, on the proof that the $\mathbb{P}_1/\mathbb{P}_1$ approximation satisfies the AGD-properties. A future work could be to prove that the $\mathbb{P}_1-\text{iso}\mathbb{P}_2/\mathbb{P}_1$ approximation also satisfies the AGD-properties.

- Finally, regarding the literature on the GDM, in [30], the authors firstly studied how a mixed mixed element method could be encompassed in the GDM framework. By the way, in this work, we have managed to ensure the H_{div} conformity of our gradient discretisation without problem dependency (contrary to [30, Remark 3.1]). Additionally we have been able here to deal with a source term in $H^{-1}(\Omega)$ and not only in $L^2(\Omega)$. Thus we have increased the cases where the mixed method can be seen as a GDM.

Chapitre 3

Energy stable and linearly well-balanced numerical schemes for the non-linear Shallow Water equations with Coriolis force

This work has been done in collaboration with Emmanuel Audusse, Noémie Gaveau and Yohan Penel. To be submitted

Contents

3.1 Introduction

The question of the accuracy of numerical schemes for hyperbolic systems with source terms around stationary solutions and/or in asymptotic regimes has been a subject of great interest over the last two decades, see the seminal works [9,36,38] in late nineties and the reference books [13,34] ten years later. In the context of geophysical flows and for colocated finite-volume methods applied to shallow water equations, a lot of works have been devoted to the accuracy around the so-called lake-at-rest equilibrium and more recently extended to nonzero velocity one dimensional stationary states, see [11] and references therein. But for large scales atmospheric or oceanographic flows, the relevant stationary state is the geostrophic equilibrium, see [53] for a general introduction to geophysical rotating fluid dynamics. The accuracy of colocated finite volume numerical schemes around such an equilibrium was less investigated, To our knowledge, the first work in this field is due to Bouchut, Le Sommer and Zeitlin [14], see also [16] and [18], but was fully accurate only for one-dimensional flows, as exhibited in [5]. Recently two independent works [42,52] proposed IMplicit-EXplicit type schemes for fully non-linear equations which are proved to be accurate near the geostrophic equilibrium but, due to their implicit part, need the solution of a global Laplacian at each time step, see also [51] for a study about the time discretization of the Coriolis term. Note that there exists also a lot of works devoted to the approximation of the Coriolis term in staggered schemes, see for example [48] for a linear analysis and [45] for the fully nonlinear case. Let us also mention the work [50] where the authors compare the dispersion relations of some mixed Finite Volume / Finite Difference and Finite Volume / Finite Element methods in a large scale oceanographic context including the Coriolis force.

In this work, we aim at designing explicit colocated finite volume schemes that are proved to be accurate around the geostrophic equilibrium and stable in the nonlinear framework. Our work is based on the ideas developed in [5] where accurate and stable Godunov type schemes were designed for the linear two-dimensional rotating wave equation but we will see in the sequel that further developments are needed to take charge of the non linear case in a conservative way. All the numerical finite volume scheme we consider in this paper belong to the AUSM family where the flux is divided in an advective part and a pressure part, see the seminal works [41] and [40] and the recent review [26]. More precisely, in Section 3.2, we first introduce the system of equations and we characterize the geostrophic equilibrium. In Section 3.3, we define some discrete operators and we prove some of their properties. Equipped with these definitions, in Section 3.4, we can define some finite volume schemes and study the two properties we are interested in : the decrease of the semi-discrete energy and the preservation of the geostrophic equilibrium in the linearized version. Moreover, in the last part of this section, we propose a conservative version of one of our scheme. Finally, in Section 3.5, we illustrate the behaviour of the schemes for some standard test cases and we exhibit a great improvement when compared to a classic finite volume scheme.

3.2 Shallow water equations and geostrophic equilibrium

Let Ω be an open bounded domain of $\mathbb{R}^2$ and let $T > 0$. The nonlinear Shallow Water equations with Coriolis force formulated on $\Omega \times (0,T)$ read :

$$\begin{cases} \partial_t h + \operatorname{div}(h\boldsymbol{u}) = 0 \,, \\ \partial_t (h\boldsymbol{u}) + \operatorname{div}(h\boldsymbol{u} \otimes \boldsymbol{u}) + h(\nabla \phi + \omega \boldsymbol{u}^\perp) = 0 \,, \end{cases} \tag{3.1}$$

where h is the water height and $\boldsymbol{u} = (u_x, u_y)$ the horizontal velocity, $\boldsymbol{u}^\perp = (-u_y, u_x)$ denoting its orthogonal vector in the (x,y) plane. The Coriolis force is accounted for in the momentum equations

through the angular speed ω. Following [20, 43], the pressure forces appear under a non conservative form through the scalar potential $\phi = gh$, where g is the standard gravity constant. For the sake of simplicity, a flat topography is considered in the present work, but the proposed approaches naturally extended to varying bottoms by taking $\phi = g(h + b)$ where b denotes the topography. It is also easy to extend the method to a varying Coriolis parameter in the β-plane approximation since all the modifications we introduce in this paper are purely local.

It is well-known that the total energy associated to the system (3.1) decomposes as $E = \mathcal{P} + \mathcal{K}$ where

$$\mathcal{P} = \frac{1}{2}gh^2 \quad \text{and} \quad \mathcal{K} = \frac{1}{2}h\|\boldsymbol{u}\|^2$$

stand respectively for potential and kinetic energies. We recall that the energy E plays the role of a mathematical entropy associated to the hyperbolic system (3.1) and regular solutions satisfy the following conservation law

$$\partial_t E + \operatorname{div}\left[\left(\phi + \frac{1}{2}\|\boldsymbol{u}\|^2\right) h\boldsymbol{u}\right] = 0\,. \tag{3.2}$$

whereas for discontinuous solution, the total energy is only non-increasing in time.

When developing numerical methods, main objectives are accuracy and stability. To get stability, a crucial objective is to build numerical approximations satisfying a discrete counterpart of (3.2) that ensures that the discrete energy is nonincreasing. To achieve this, a general strategy is to consider a sufficient amount of numerical diffusion in the scheme. But in some physical contexts such as low Froude number regimes or near specific stationary states, these diffusive terms may considerably degrade the accuracy of the approximations and specific schemes are needed. Here we are interested in flows around the geostrophic balance

$$\nabla\phi + \omega\boldsymbol{u}^\perp = 0, \quad \operatorname{div}\boldsymbol{u} = 0 \tag{3.3}$$

To address such an issue, based on the study for the linear case [5], we propose a numerical approach involving discrete versions of these equilibria in the numerical fluxes. As a preliminary step, the strategy can be understood at the continuous level by investigating how the model (3.1) behaves with respect to some generic perturbations $(\boldsymbol{q}, \pi)$:

$$\begin{cases} \partial_t h + \operatorname{div}(h\boldsymbol{u} - \boldsymbol{q}) = 0\,, \\ \partial_t(h\boldsymbol{u}) + \operatorname{div}\big(\boldsymbol{u} \otimes (h\boldsymbol{u} - \boldsymbol{q})\big) + (h\nabla\phi - \nabla\pi) + \omega\,(h\boldsymbol{u} - \boldsymbol{q})^\perp = 0\,. \end{cases} \tag{3.4}$$

or, in conservative form

$$\begin{cases} \partial_t h + \operatorname{div}(h\boldsymbol{u} - \boldsymbol{q}) = 0 \\ \partial_t(h\boldsymbol{u}) + \operatorname{div}\big(\boldsymbol{u} \otimes (h\boldsymbol{u} - \boldsymbol{q})\big) + \nabla\left(\frac{1}{2}gh^2 - \pi\right) + \omega\,(h\boldsymbol{u} - \boldsymbol{q})^\perp = 0\,. \end{cases} \tag{3.5}$$

where $\boldsymbol{q}$ and π can be respectively seen as (small) perturbations with respect to the flow rate and to the hydrostatic pressure.

The solutions to the modified equations (3.4) satisfy the following energy balance :

$$\partial_t E + \operatorname{div}\left[\left(\phi + \frac{1}{2}\|\boldsymbol{u}\|^2\right)\ (h\boldsymbol{u} - \boldsymbol{q}) - \pi\boldsymbol{u}\right] = -\boldsymbol{q}\cdot\left(\nabla\phi + \omega\boldsymbol{u}^\perp\right) - \pi\ \operatorname{div}\boldsymbol{u}\,, \tag{3.6}$$

which motivates a choice for $\boldsymbol{q}$ and π involving resp. the quantities $\nabla\phi + \omega\,\boldsymbol{u}^\perp$ and $\operatorname{div}\boldsymbol{u}$. Let us remark that these quantities govern the geostrophic equilibrium (3.3) associated to System (3.1) linearized around the steady state $(\tilde{h}, \widetilde{\boldsymbol{u}}) = (h_0, 0)$ for a constant h_0 :

$$\begin{cases} \partial_t h = -h_0\ \operatorname{div}\boldsymbol{u}\ , \\ \partial_t\boldsymbol{u} = -(\nabla\phi + \omega\,\boldsymbol{u}^\perp)\,. \end{cases}$$

From a numerical point of view, diffusion terms are thus expected to have regularizing effects in the sense that they allow to recover a discrete counterpart of (3.6). Moreover, such terms are intended to vanish close to the geostrophic equilibrium, which must improve the quality of the approximations in this regime. The rest of the article is then devoted to the presentation of different ways to implement this idea in a discrete setting.

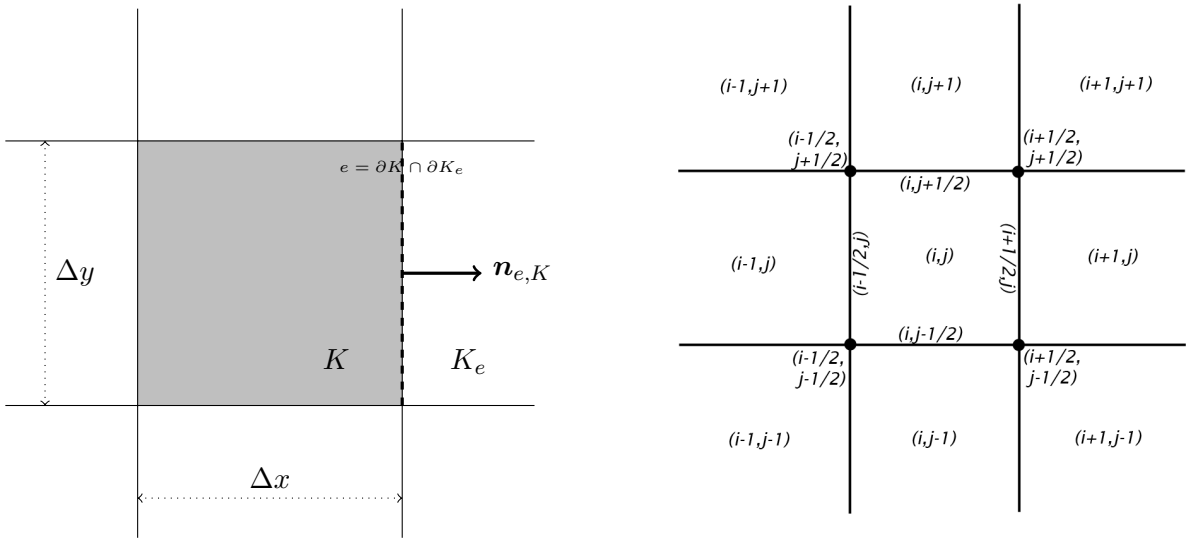
3.3 Discrete operators

3.3.1 Definition of the mesh

Let us first introduce some generic notations related to the discretization of the equations. We consider a tessellation $\mathbb{T}$ of the computational domain $\Omega \subset \mathbb{R}^2$ made of non-overlapping rectangular cells of sizes Δx, Δy. The set of all edges of the mesh is denoted by $\mathcal{E}$ and the set of vertices by $\mathcal{V}$.

- A generic cell of $\mathbb{T}$ is denoted by K, its area by m_K and its boundary by ∂K. A given quantity Φ located on K is numbered $\Phi_{i,j}$.
- A generic edge of $\mathcal{E}$ is denoted by e and its length by m_e. A given quantity Φ located on e is numbered $\Phi_{i+1/2,j}$ (respectively $\Phi_{i,j+1/2}$) for y-axis (respectively x-axis) edge.
- Given a cell K and an edge $e \in \partial K$, K_e is the neighbouring cell to e (other than K) and $\boldsymbol{n}_{e,K}$ the outward normal pointing to K_e.
- A generic vertex of $\mathcal{V}$ is denoted by v. A given quantity Φ located on v is numbered $\Phi_{i+1/2,j+1/2}$.

Notations are pictured on Figs. 3.1.



(a) Mesh notations

(b) Variables numbering

FIGURE 3.1 – Geometric notations

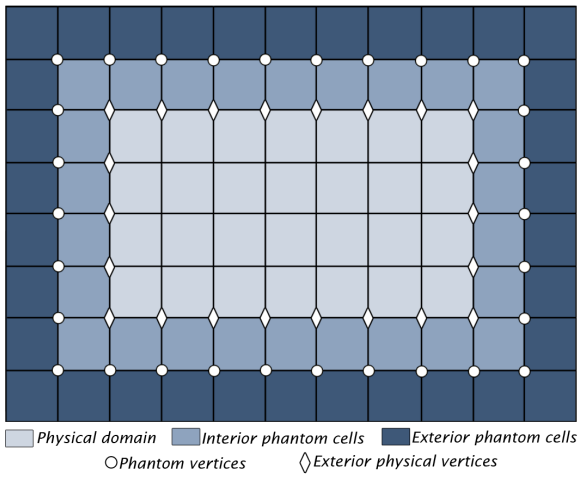


FIGURE 3.2 – Physical and phantom domains

3.3.2 Discrete operators

Equipped with these geometrical settings, we can now introduce discrete operators that will be needed to construct numerical schemes. Since we only consider colocated finite volume schemes, all the unknowns are defined on the cells $K \in \mathbb{T}$. But we will see in the next sections that some other quantities (including the numerical diffusion terms) need to be computed on the edges $e \in \mathcal{E}$ or at the vertices $v \in \mathcal{V}$. Then we need to define discrete operators from cells to edges (and vice-versa) and from cells to vertices (and vice-versa) – see Figure 3.1. In the following definitions, the notations $X_i^j(\phi_j)$ means that the operator X is applied to a quantity ϕ defined at the location j and computes a quantity that is defined at the location i. For example, the first operator below $\nabla_e^K \varphi_K$ is a discrete gradient operator that is defined for quantities that are defined on a cell K and that allows to construct a consistent gradient on an edge e. Discrete gradient and divergence are denoted with classic notations. The notation f always denotes an algebraic reconstruction operator. Let us begin with the operators from cells to edges (and vice-versa).

$$\nabla_e^K \varphi_K = \frac{m_e}{m_K}(\varphi_{Ke} - \varphi_K)\ \boldsymbol{n}_{e,K} \qquad \nabla_K^e \varphi_e = \frac{1}{m_K}\sum_{e\subset\partial K} m_e\, \varphi_e\, \boldsymbol{n}_{e,K}$$

$$\mathrm{div}_e^K \boldsymbol{\varphi}_K = \frac{m_e}{m_K}(\boldsymbol{\varphi}_{Ke} - \boldsymbol{\varphi}_K)\cdot \boldsymbol{n}_{e,K} \qquad \mathrm{div}_K^e \boldsymbol{\varphi_e} = \frac{1}{m_K}\sum_{e\subset\partial K} m_e\, \boldsymbol{\varphi}_e\cdot \boldsymbol{n}_{e,K}$$

$$f_e^K(\varphi_K) = \frac{1}{2}(\varphi_{Ke} + \varphi_K) \qquad f_K^e(\boldsymbol{\varphi}_e) = \frac{1}{2}\sum_{e\subset\partial K} \boldsymbol{\varphi}_e\cdot \boldsymbol{n}_{e,K}\ \boldsymbol{n}_{e,K}$$

$$f_e^K(\boldsymbol{\varphi}_K) = \frac{1}{2}(\boldsymbol{\varphi}_{Ke} + \boldsymbol{\varphi}_K)\cdot \boldsymbol{n}_{e,K}\ \boldsymbol{n}_{e,K}$$

Then we define operators from cells to vertices, where we use the notation $\boldsymbol{\varphi} = (\varphi, \psi)$

$$[\nabla_v^K \varphi_K]_{i+1/2,j+1/2} = \frac{1}{2}\begin{pmatrix} \dfrac{\varphi_{i+1,j+1} - \varphi_{i,j+1}}{\Delta x} + \dfrac{\varphi_{i+1,j} - \varphi_{i,j}}{\Delta x} \\ \dfrac{\varphi_{i+1,j+1} - \varphi_{i+1,j}}{\Delta y} + \dfrac{\varphi_{i,j+1} - \varphi_{i,j}}{\Delta y} \end{pmatrix}$$

$$[\mathrm{div}_v^K \boldsymbol{\varphi}_K]_{i+1/2,j+1/2} = \frac{1}{2}\left[\frac{\varphi_{i+1,j+1} - \varphi_{i,j+1}}{\Delta x} + \frac{\varphi_{i+1,j} - \varphi_{i,j}}{\Delta x}\right] + \frac{1}{2}\left[\frac{\psi_{i+1,j+1} - \psi_{i+1,j}}{\Delta y} + \frac{\psi_{i,j+1} - \psi_{i,j}}{\Delta y}\right]$$

$$[f_v^K(\boldsymbol{\varphi}_K)]_{i+1/2,j+1/2} = \frac{\boldsymbol{\varphi}_{i+1,j+1} + \boldsymbol{\varphi}_{i,j+1} + \boldsymbol{\varphi}_{i+1,j} + \boldsymbol{\varphi}_{i,j}}{4}$$

and from vertices to cells

$$
\begin{aligned}
[\nabla_K^v \varphi_v]_{i,j} &= \frac{1}{2}\begin{pmatrix} \dfrac{\varphi_{i+1/2,j+1/2}-\varphi_{i-1/2,j+1/2}}{\Delta x} + \dfrac{\varphi_{i+1/2,j-1/2}-\varphi_{i-1/2,j-1/2}}{\Delta x} \\ \dfrac{\varphi_{i+1/2,j+1/2}-\varphi_{i+1/2,j-1/2}}{\Delta y} + \dfrac{\varphi_{i-1/2,j+1/2}-\varphi_{i-1/2,j-1/2}}{\Delta y} \end{pmatrix} \\
[\mathrm{div}_K^v \boldsymbol{\varphi}_v]_{i,j} &= \frac{1}{2}\left[\frac{\varphi_{i+1/2,j+1/2}-\varphi_{i-1/2,j+1/2}}{\Delta x} + \frac{\varphi_{i+1/2,j-1/2}-\varphi_{i-1/2,j-1/2}}{\Delta x}\right] \\
&\quad + \frac{1}{2}\left[\frac{\psi_{i+1/2,j+1/2}-\psi_{i+1/2,j-1/2}}{\Delta y} + \frac{\psi_{i-1/2,j+1/2}-\psi_{i-1/2,j-1/2}}{\Delta y}\right] \\
[f_K^v(\boldsymbol{\varphi}_v)]_{i,j} &= \frac{\boldsymbol{\varphi}_{i+1/2,j+1/2}+\boldsymbol{\varphi}_{i-1/2,j+1/2}+\boldsymbol{\varphi}_{i+1/2,j-1/2}+\boldsymbol{\varphi}_{i-1/2,j-1/2}}{4}
\end{aligned}
$$

We will also need a divergence operator from edges to vertices

$$
[\mathrm{div}_v^e \boldsymbol{\Phi}_e]_{i+1/2,j+1/2} = \frac{1}{\Delta x}\left(\boldsymbol{\Phi}_{i+1,j+1/2}-\boldsymbol{\Phi}_{i,j+1/2}\right)\cdot \boldsymbol{e}_x + \frac{1}{\Delta y}\left(\boldsymbol{\Phi}_{i+1/2,j+1}-\boldsymbol{\Phi}_{i+1/2,j}\right)\cdot \boldsymbol{e}_y.
$$

Finally we need to define upwind divergence operators that will be used to discretize the transport part of the equations in order to ensure the stability of the numerical schemes. In the following definitions, the quantity

$$
\varphi^{\pm} = \frac{1}{2}(\varphi \pm |\varphi|)
$$

will refer to the positive and negative parts of any scalar function φ. From edges to cells, the operator reads

$$
\mathbf{div}_K^{e,up}(\boldsymbol{\psi}_K \otimes \boldsymbol{\varphi}_e) = \frac{1}{m_K}\sum_{e\subset\partial K} m_e\left(\boldsymbol{\psi}_K(\boldsymbol{\varphi}_e\cdot\boldsymbol{n}_{e,K})^+ + \boldsymbol{\psi}_{K_e}(\boldsymbol{\varphi}_e\cdot\boldsymbol{n}_{e,K})^-\right)
$$

and from vertices to cells, it reads

$$
\begin{aligned}
&\mathbf{div}_K^{v,up}(\boldsymbol{\psi}_K \otimes \boldsymbol{\varphi}_v)_{i,j} \\
&= \frac{1}{2\Delta x}\Big(\boldsymbol{\psi}_{i,j}\left((\boldsymbol{\varphi}_{i+1/2,j+1/2}+\boldsymbol{\varphi}_{i+1/2,j-1/2})\cdot\boldsymbol{e}_x\right)^+ + \boldsymbol{\psi}_{i+1,j}\left((\boldsymbol{\varphi}_{i+1/2,j+1/2}+\boldsymbol{\varphi}_{i+1/2,j-1/2})\cdot\boldsymbol{e}_x\right)^- \\
&\quad + \boldsymbol{\psi}_{i,j}\left((\boldsymbol{\varphi}_{i-1/2,j+1/2}+\boldsymbol{\varphi}_{i-1/2,j-1/2})\cdot(-\boldsymbol{e}_x)\right)^+ + \boldsymbol{\psi}_{i-1,j}\left((\boldsymbol{\varphi}_{i-1/2,j+1/2}+\boldsymbol{\varphi}_{i-1/2,j-1/2})\cdot(-\boldsymbol{e}_x)\right)^-\Big) \\
&\quad + \frac{1}{2\Delta y}\Big(\boldsymbol{\psi}_{i,j}\left((\boldsymbol{\varphi}_{i+1/2,j+1/2}+\boldsymbol{\varphi}_{i-1/2,j+1/2})\cdot\boldsymbol{e}_y\right)^+ + \boldsymbol{\psi}_{i,j+1}\left((\boldsymbol{\varphi}_{i+1/2,j+1/2}+\boldsymbol{\varphi}_{i-1/2,j+1/2})\cdot\boldsymbol{e}_y\right)^- \\
&\quad + \boldsymbol{\psi}_{i,j}\left((\boldsymbol{\varphi}_{i+1/2,j-1/2}+\boldsymbol{\varphi}_{i-1/2,j-1/2})\cdot(-\boldsymbol{e}_y)\right)^+ + \boldsymbol{\psi}_{i,j-1}\left((\boldsymbol{\varphi}_{i+1/2,j-1/2}+\boldsymbol{\varphi}_{i-1/2,j-1/2})\cdot(-\boldsymbol{e}_y)\right)^-\Big).
\end{aligned}
$$

3.3.3 Mimetic properties

These discrete operators satisfy some important properties that will be used to prove some results for the numerical schemes we propose in the next section.

Lemma 3.3.1

We first mention some local properties about the permutation of the derivative, reconstruction and orthogonal operators. Computations are obvious.

i) $f_v^K(\boldsymbol{u}_K^\perp) = (f_v^K\,\boldsymbol{u}_K)^\perp$ *and* $f_K^v(\boldsymbol{u}_K^\perp) = (f_K^v\,\boldsymbol{u}_K)^\perp$

ii) $-\operatorname{div}_v^e((f_e^K(\boldsymbol{u}_K^\perp))^\perp) = \operatorname{div}_v^K \boldsymbol{u}_K$

iii) $\operatorname{div}_K^v(f_v^K\,\boldsymbol{u}_K) = f_K^v(\operatorname{div}_v^K \boldsymbol{u}_K)$

We now define the three scalar products

$$\langle \boldsymbol{\Phi}_K, \boldsymbol{\Psi}_K\rangle = \sum_{K\in\mathbb{T}} \boldsymbol{\Phi}_K\cdot\boldsymbol{\Psi}_K, \quad \langle \boldsymbol{\Phi}_v, \boldsymbol{\Psi}_v\rangle = \sum_{v\in\mathcal{V}} \boldsymbol{\Phi}_v\cdot\boldsymbol{\Psi}_v$$

$$\text{and} \quad \langle \boldsymbol{\varphi}_e, \boldsymbol{\Psi}_e\rangle = \sum_{i,j}\left[(\boldsymbol{\varphi}_{i+1/2,j}\cdot\boldsymbol{e}_x)(\boldsymbol{\Psi}_{i+1/2,j}\cdot\boldsymbol{e}_x) + (\boldsymbol{\varphi}_{i,j+1/2}\cdot\boldsymbol{e}_y)(\boldsymbol{\Psi}_{i,j+1/2}\cdot\boldsymbol{e}_y)\right].$$

Lemma 3.3.2

We have the following properties for the reconstruction and orthogonal operators

i) $\langle \boldsymbol{u}_K, (f_K^e\,\boldsymbol{q}_e)^\perp\rangle = -\langle f_e^K(\boldsymbol{u}_K^\perp), \boldsymbol{q}_e\rangle$

ii) $\langle \boldsymbol{u}_K, f_K^v(\boldsymbol{q}_v^\perp)\rangle = -\langle f_v^K(\boldsymbol{u}_K^\perp), \boldsymbol{q}_v\rangle$

iii) $\langle f_K^v(f_v^K(\boldsymbol{u}_K^\perp)), \boldsymbol{u}_K\rangle = 0$

iv) $\langle f_K^v(f_v^K(h_k)\,f_v^K(\boldsymbol{u}_K^\perp)), \boldsymbol{u}_K\rangle = 0$

Proof: Property i) is obtained after a rearrangement of sum under the condition $\boldsymbol{q}_e = 0$ on the edges on the boundary of the physical domain (see Figure 3.2) to eliminate the boundary terms. Property ii) is obtained after a rearrangement of sum under the periodic boundary conditions on h_K and $\boldsymbol{u}_K$ and $\boldsymbol{q}_v = 0$ on the vertices on the boundary of the interior domain (see Figure 3.2) to eliminate the boundary terms.
Property iii) and lem :reconstruction :vii are obtained after a rearrangement of sum under the periodic boundary conditions on h_K and $\boldsymbol{u}_K$ to eliminate the boundary terms. ■

Lemma 3.3.3

We have the following mimetic properties for the discrete gradient and divergence operators, including for some of them the reconstruction operators

i) $\langle \phi_K, \mathrm{div}_K^e \, \boldsymbol{q}_e \rangle = -\langle \nabla_e^K \, \phi_K, \boldsymbol{q}_e \rangle$

ii) $\langle \pi_v, \mathrm{div}_v^K \, \boldsymbol{u}_K \rangle = -\langle \nabla_K^v \, \pi_v, \boldsymbol{u}_K \rangle$

iii) $\langle \phi_K, \mathrm{div}_K^v \, \boldsymbol{q}_v \rangle = -\langle \nabla_v^K \, \phi_K, \boldsymbol{q}_v \rangle$

iv) $\langle f_K^e(\nabla_e^K \, \phi_K), h_K \boldsymbol{u}_K \rangle = -\langle \phi_K, \mathrm{div}_K^e(f_e^K(h_K \boldsymbol{u}_K)) \rangle$

v) $\langle f_K^v(\nabla_v^K \, \phi_K), \boldsymbol{u}_K \rangle = -\langle \phi_K, f_K^v(\mathrm{div}_v^K \, \boldsymbol{u}_K) \rangle$

vi) $\langle f_K^v(f_v^K(h_K) \, \nabla_v^K \, \phi_K), \boldsymbol{u}_K \rangle = -\langle \phi_K, \mathrm{div}_K^v(f_v^K(h_K) \, f_v^K(\boldsymbol{u}_K)) \rangle$

Proof: Property i) (respectively ii)) is obtained after a rearrangement of sum under the condition $\boldsymbol{q}_e = 0$ on the edges (respectively $\pi_v = 0$ on the vertices) on the boundary of the physical domain (see Figure 3.2) to eliminate the boundary terms.
Property iii) is obtained after a rearrangement of sum under the periodic boundary conditions on h_K and $\boldsymbol{u}_K$ and $\boldsymbol{q}_v = 0$ on the vertices on the boundary of the interior domain (see Figure 3.2) to eliminate the boundary terms.
Property iv), v) and vi) are obtained after a rearrangement of sum under the periodic boundary conditions on h_K and $\boldsymbol{u}_K$ to eliminate the boundary terms. ■

3.4 Well-balanced and stable finite volume schemes

When considering hyperbolic equations, classic finite volume schemes in colocated two-dimensional cartesian framework are often referred as five points schemes. Indeed the update of the quantities of interest in a cell K of the tesselation $\mathbb{T}$ needs the computation of the fluxes through the four edges of its boundary ∂K, see Figure 3.1. For a first order scheme, the numerical flux through an edge $e \in \mathcal{E}$ is generally computed from the values of the quantities in the two neighbouring cells. Hence, the update of the quantities in a cell K involves five cells of the tesselation $\mathbb{T}$. We refer to [39, 49] for more details about classic first order finite volume schemes for hyperbolic problems. In the last section of the paper, devoted to the numerical tests, such a five point scheme, typically the HLLC scheme, see [49], will be considered as a standard scheme to which we will compare the well-balanced schemes we designed in the next sections. Because of the essentially two-dimensional character of the geostrophic equilibrium, these well-balanced schemes have to involve a larger stencil. Note that to consider enlarged stencils is also a common way to design high order schemes in a finite volume framework through the MUSCL strategy, see [39, 49], and is also commonly used in diffusion problems since the computation of the numerical flux needs the reconstruction of a complete two-dimensional gradient, see [29].

3.4.1 Edge based 9 points schemes

In this section, we present two nine points schemes for which the geostrophic equilibrium (3.3) and the perturbation q, see (3.4), are defined at the edges e whereas the perturbation π is defined on the vertices v. None of them succeeds to ensure both the preservation of the linearized stationary states

and the decreasing of the nonlinear energy. nevertheless, the numerical results presented in the last section of the paper exhibit a reasonably good behaviour for all the test cases we have performed.

Energy decreasing semi-discrete scheme

In [6], we proposed a first numerical scheme for which we prove a semi-discrete counterpart of (3.6), i.e. the semi-discrete energy is non-increasing. Note that all along the paper the term semi-discrete will refer to quantities that are discrete in space but continuous in time. We also exhibited in that work that the linearized version of the scheme failed to preserve the geostrophic balance. Using the discrete operators introduced in Section 3.3, the scheme now reads

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{dt}} h_K + \mathrm{div}_K^e(\boldsymbol{\mathcal{F}}_e) = 0, \\ \dfrac{\mathrm{d}}{\mathrm{dt}} (h_K \boldsymbol{u}_K) + \mathbf{div}_K^{e,up}(\boldsymbol{u}_K \otimes \boldsymbol{\mathcal{F}}_e) + h_K \; f_K^e(\nabla_e^K \, \phi_K) - \nabla_K^v \, \pi_v = -\omega \, (\, h_K \boldsymbol{u}_K - f_K^e \, \boldsymbol{q}_e)^\perp \, , \end{cases} \tag{3.7}$$

where the mass fluxes are defined at the level of the edge e as :

$$\boldsymbol{\mathcal{F}}_e = f_e^K(h_K \boldsymbol{u}_K) - \boldsymbol{q}_e \, , \tag{3.8}$$

with the numerical diffusion term on the flow rate

$$\boldsymbol{q}_e = \gamma \, \frac{\Lambda}{g} \max\{\Delta x, \Delta y\} (\omega \, f_e^K(\boldsymbol{u}_K^\perp) + \nabla_e^K \, \phi_K), \tag{3.9}$$

while the numerical diffusion term on the hydrostatic pressure is defined at the vertices

$$\pi_v = \nu \Lambda \, \max\{\Delta x, \Delta y\} \, f_v^K(h_K) \, \mathrm{div}_v^K(\boldsymbol{u}_K), \tag{3.10}$$

where γ and ν are positive dimensionless constants, and Λ is a positive characteristic velocity, typically, we take

$$\Lambda = \max_{K \in \mathbb{T}} \left\{ \|\boldsymbol{u}_K\| + \sqrt{g h_K} \right\}.$$

Semi-discrete energy : We first show the scheme (3.7) ensures a discrete counterpart of (3.6) through semi-discrete mechanic energy estimates. To do so, we first need the two following lemmas, describing the evolution of potential and kinetic energies.

Lemma 3.4.1 (*Semi-discrete potential energy*)

We set $\mathcal{P}_K = \dfrac{1}{2} g \, (h_K)^2$ *for* $K \in \mathbb{T}$. *Then :*

$$\frac{\mathrm{d}}{\mathrm{dt}} \mathcal{P}_K + \phi_K \, \mathrm{div}_K^e(f_e^K(h_K \boldsymbol{u}_K)) - \phi_K \, \mathrm{div}_K^e \, \boldsymbol{q}_e = 0. \tag{3.11}$$

Lemma 3.4.2 (*Semi-discrete kinetic energy*)

We set $\mathcal{K}_K = \dfrac{1}{2} h_K \|\boldsymbol{u}_K\|^2$ *for* $K \in \mathbb{T}$. *Then :*

$$\frac{\mathrm{d}}{\mathrm{dt}} \mathcal{K}_K + \frac{1}{2} \, \mathbf{div}_K^{e,up} \left(\|\boldsymbol{u}_K\|^2 \boldsymbol{\mathcal{F}}_e \right) + h_K \boldsymbol{u}_K \cdot f_K^e \left(\nabla_e^K \, \phi_K \right) \leq \omega \boldsymbol{u}_K \cdot (f_K^e \, \boldsymbol{q}_e)^\perp + \boldsymbol{u}_K \cdot \nabla_K^v \, \pi_v. \tag{3.12}$$

Proof: We write :

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}\left(\frac{1}{2}h_K\|\boldsymbol{u}_K\|^2\right) &= \boldsymbol{u}_K\cdot\frac{\mathrm{d}}{\mathrm{dt}}(h_K\boldsymbol{u}_K)-\frac{1}{2}\|\boldsymbol{u}_K\|^2\frac{\mathrm{d}}{\mathrm{dt}}h_K\\ &= -h_K\boldsymbol{u}_K\cdot f_K^e(\nabla_e^K\,\phi_K)-\boldsymbol{u}_K\cdot\mathbf{div}_K^{e,up}\left(\boldsymbol{u}_K\otimes\boldsymbol{\mathcal{F}}_e\right)+\omega\boldsymbol{u}_K\cdot(f_K^e\,\boldsymbol{q}_e)^\perp)\\ &\quad+\frac{1}{2}\|\boldsymbol{u}_K\|^2\,\mathrm{div}_K^e\,\boldsymbol{\mathcal{F}}_e+\boldsymbol{u}_K\cdot\nabla_K^v\,\pi_v.\end{aligned}$$

After some basic computations, we get the relation

$$\begin{aligned}&\frac{1}{2}\|\boldsymbol{u}_K\|^2\,\mathrm{div}_K^e\,\boldsymbol{\mathcal{F}}_e-\boldsymbol{u}_K\cdot\mathbf{div}_K^{e,up}\left(\boldsymbol{u}_K\otimes\boldsymbol{\mathcal{F}}_e\right)=\\ &\qquad-\frac{1}{2}\,\mathbf{div}_K^{e,up}(\|\boldsymbol{u}_K\|^2\boldsymbol{\mathcal{F}}_e)+\frac{1}{2m_K}\sum_{e\subset\partial K}\|\boldsymbol{u}_K-\boldsymbol{u}_{Ke}\|^2\left(\boldsymbol{\mathcal{F}}_e\cdot\boldsymbol{n}_{e,k}\right)^-\ .\end{aligned}\tag{3.13}$$

The second term of the right hand side being non-positive, we get the announced result . ■

Proposition 3.4.3 (*Decreasing of the semi-discrete energy*)

We define the total energy $E_K=\mathcal{P}_K+\mathcal{K}_K$. Then we obtain a discrete counterpart of (3.6)

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{dt}}\left(\sum_{K\in\mathbb{T}}m_KE_K\right)\\ &\qquad\leq-\max\{\Delta x,\Delta y\}\sum_{K\in\mathbb{T}}m_K\left[\gamma\,\frac{\Lambda}{g}\|\omega\,f_e^K(\boldsymbol{u}_K^\perp)+\nabla_e^K\,\phi_K\|^2+\nu\Lambda\ f_v^K(h_K)(\mathrm{div}_v^K\,\boldsymbol{u}_K)^2\right].\end{aligned}$$

Proof: Gathering relations (3.11) and (3.12), we obtain the following estimate for the total energy

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{dt}}E_K+\frac{1}{2}\,\mathbf{div}_K^{e,up}\left(\|\boldsymbol{u}_K\|^2\boldsymbol{\mathcal{F}}_e\right)+\phi_K\,\mathrm{div}_K^e(f_e^K(h_K\boldsymbol{u}_K))+h_K\boldsymbol{u}_K\cdot f_K^e\left(\nabla_e^K\,\phi_K\right)\\ &\qquad\leq-\boldsymbol{u}_K\cdot(f_K^e\,\boldsymbol{q}_e)^\perp+\phi_K\,\mathrm{div}_K^e\,\boldsymbol{q}_e\ +\boldsymbol{u}_K\cdot\nabla_K^v\,\pi_v.\end{aligned}$$

Hence we get the following semi-discrete inequality using Lemma 3.3.2 and 3.3.3

$$\frac{\mathrm{d}}{\mathrm{dt}}\left(\sum_{K\in\mathbb{T}}m_KE_K\right)\leq\sum_{K\in\mathbb{T}}m_K\left[-\boldsymbol{q}_e\cdot(\omega\,f_e^K(\boldsymbol{u}_K^\perp)+\nabla_e^K\,\phi_K)-\pi_v\,\mathrm{div}_v^K\,\boldsymbol{u}_K\right].$$

The results follows from the choices of the numerical diffusion terms (3.9) and (3.10). ■

Linearized well-balanced property : We refer to the next paragraph for the details of the proof. Here it is obvious that the linearized version of the scheme can not exactly preserve the geostrophic equilibrium. Indeed the numerical diffusion term $\boldsymbol{q}_e\propto\omega\,f_e^K(\boldsymbol{u}_K^\perp)+\nabla_e^K\,\phi_K$ (3.9) is defined on the edges whereas the geostrophic term $\omega\boldsymbol{u}_K^\perp+f_K^e(\nabla_e^K\,\phi_K)$ that appears in the update of the momentum in (3.7) is defined on the cells. It follows that if we define the discrete geostrophic equilibrium on the edges, the numerical diffusion term will vanish but not the other one, and vice versa. This is due to the chosen operators which ensure $f_K^e(f_e^K(\boldsymbol{u}_K^\perp))\neq\boldsymbol{u}_K^\perp$.

Well-balanced semi-discrete scheme

For completeness, we propose here a slightly modified scheme whose linearized version is meant to preserve the geostrophic equilibrium but for which it is not possible to ensure that the semi-discrete energy is non-increasing. This semi-discrete scheme reads

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{dt}} h_K + \mathrm{div}_K^e \, \boldsymbol{\mathcal{F}}_e = 0, \\ \dfrac{\mathrm{d}}{\mathrm{dt}} (h_K \boldsymbol{u}_K) + \mathbf{div}_K^{e,up}(\boldsymbol{u}_K \otimes \boldsymbol{\mathcal{F}}_e) + h_K \, f_K^e(\nabla_e^K \, \phi_K) - \nabla_K^v \, \pi_v = -\omega h_K \, f_K^e(f_e^K(\boldsymbol{u}_K^\perp)) + \omega (f_K^e \, \boldsymbol{q}_e)^\perp, \end{cases} \tag{3.14}$$

with definitions (3.8), (3.9) and (3.10) for the mass flux, the numerical correction on the flow rate and the numerical correction on the pressure term. We see that only the definition of the Coriolis term differs from the previous version.

Linearized well-balanced property : The linearized version of the scheme (3.14) reads

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{dt}} h_K + h_0 \ \mathrm{div}_K^e(f_e^K(\boldsymbol{u}_K)) - h_0 \, \mathrm{div}_K^e \, \tilde{\boldsymbol{q}}_e = 0, \\ \dfrac{\mathrm{d}}{\mathrm{dt}} \boldsymbol{u}_K + \ f_K^e(\nabla_e^K \, \phi_K) - \ \nabla_K^v \, \tilde{\pi}_v = -\omega \, f_K^e(f_e^K(\boldsymbol{u}_K^\perp)) + \omega (f_K^e \, \tilde{\boldsymbol{q}}_e)^\perp \, , \end{cases} \tag{3.15}$$

with the numerical diffusion terms

$$\tilde{\boldsymbol{q}}_e = \gamma \, \frac{\Lambda}{g h_0} \max\{\Delta x, \Delta y\} \, \left(\omega \, f_e^K(\boldsymbol{u}_K^\perp) + \nabla_e^K \, \phi_K \right) \quad \text{and} \quad \tilde{\pi}_v = \nu \Lambda \, \max\{\Delta x, \Delta y\} \, \mathrm{div}_v^K \, \boldsymbol{u}_K.$$

Proposition 3.4.4

Without considering the term $\mathrm{div}_K^e(f_e^K \, \boldsymbol{u}_K)$*, the linearized scheme* (3.15) *preserves the discrete geostrophic equilibrium* $\tilde{\boldsymbol{q}}_e = 0$*.*

Proof: When the geostrophic balance expressed on the edges holds, i.e when $\omega \, f_e^K(\boldsymbol{u}_K^\perp) + \nabla_e^K \, \phi_K = 0$, the geostrophic term that appears in the update of the momentum obviously vanishes

$$f_K^e(\nabla_e^K \, \phi_K) + \omega \, f_K^e(f_e^K(\boldsymbol{u}_K^\perp)) = f_K^e \left(\nabla_e^K \, \phi_K + \omega \, f_e^K(\boldsymbol{u}_K^\perp) \right) = 0.$$

Moreover, thanks to Lemma 3.3.1, we have

$$\mathrm{div}_v^K \, \boldsymbol{u}_K = -\, \mathrm{div}_v^e((f_e^K(\boldsymbol{u}_K^\perp))^\perp) = \mathrm{div}_v^e \left(\frac{(\nabla_e^K \, \phi_K)^\perp}{\omega} \right)$$

and obvious computations show that $\mathrm{div}_v^e \left((\nabla_e^K \, \phi_K)^\perp \right) = 0$. ■

Remark 3.4.5

The term $\mathrm{div}_K^e(f_e^K \, \boldsymbol{u}_K)$ *is not strictly equal to zero when the geostrophic equilibrium holds on the edges of the considered cell. However, by enforcing the geostrophic equilibrium on enough edges of the domain it is possible to write this term on the cell* (i, j) *using solely velocities on cells* $(i \pm 2$ *and/or* $j \pm 2)$*. Due to the complexity of the computations, we failed to prove by recurrence that this term on the cell* (i, j) *can be written using solely velocities on cells* $(i \pm k$ *and/or* $j \pm k)$*, with* k *arbitrarily large. Nevertheless, the test cases presented in Section 3.5 tend to show that this term is not major and that scheme* (3.14) *has similar results to other well-balanced scheme. Hence, we assume in the following that scheme* (3.14) *can be considered as a well-balanced scheme and will be referred to as such.*

Remark 3.4.6

As it contains a reconstruction operator on the velocity, the discrete geostrophic equilibrium $\tilde{\boldsymbol{q}}_e = 0$ *also contains spurious solution*

$$\forall (i,j), \quad \phi_{i,j} = cst \quad and \quad \boldsymbol{u}_{i,j} = \left((-1)^i\, a_j, (-1)^j\, b_i\right)^T.$$

Due to the nonlinear term on the velocity, this spurious solution does not belong to the kernel of the nonlinear scheme (3.14)*. However, when* $\forall j,\ a_j = a$ *and* $\forall i,\ b_i = b$*, this solution does appear in the kernel of the nonlinear scheme. Numerically, these solutions do not appear in the test cases presented in Section 3.5. One can note that in the edge-based entropic scheme* (3.7)*, there is no reconstruction operator on the velocity and it is difficult to formulate a spurious solution due to the fact that such solution has to verify :*

$$\begin{cases} \omega\, f_e^K(\boldsymbol{u}_K^\perp) + \nabla_e^K\, \phi_K = & 0 \\ \omega \boldsymbol{u}_K^\perp + f_K^e(\nabla_e^K\, \phi_K) = & 0. \end{cases}$$

Numerically, simulations in Section 3.5 do not show such solutions.

Semi-discrete energy : The computations are similar to the previous scheme but the Coriolis term is no more the same and some computations show that

$$\langle f_K^e\left(f_e^K(\boldsymbol{u}_K^\perp)\right), \boldsymbol{u}_K\rangle = \frac{1}{2}\sum_{i,j}\left(v_{i,j}\frac{u_{i,j+1}+u_{i,j-1}}{2} - u_{i,j}\frac{v_{i+1,j}+v_{i-1,j}}{2}\right).$$

Hence, it remains a term with no *a priori* sign on the right hand side of the semi-discrete energy inequality. Note that numerical simulations presented in Section 3.5 tend to show that this term is usually negligible with respect to the non positive correction terms.

3.4.2 Cell-based 13 points scheme

In this section, we slightly enlarged the stencil in order to propose a scheme for which we are able to prove that the semi-discrete energy is non-increasing and that the linearized version does preserve the geostrophic equilibrium. To do so, we define the geostrophic equilibrium and the perturbation q on the cells K and the perturbation π on the vertices v and finally obtain a scheme with a thirteen points stencil.

The semi-discrete scheme reads

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{dt}} h_K + \mathrm{div}_K^e(f_e^K\, \boldsymbol{\mathcal{F}}_K) = 0, \\ \dfrac{\mathrm{d}}{\mathrm{dt}}(h_K \boldsymbol{u}_K) + \mathbf{div}_K^{e,up}(\boldsymbol{u}_K \otimes f_e^K\, \boldsymbol{\mathcal{F}}_K) + h_K\, \nabla_K^e(f_e^K\, \phi_K) - \nabla_K^v\, \pi_v = -\omega\,(\, h_K \boldsymbol{u}_K - \boldsymbol{q}_K)^\perp, \end{cases} \tag{3.16}$$

where the mass flux $f_e^K\, \boldsymbol{\mathcal{F}}_K$ is defined on the edge e thanks to a reconstruction operator applied to a modified momentum term initially defined on the cells K by :

$$\boldsymbol{\mathcal{F}}_K = h_K \boldsymbol{u}_K - \boldsymbol{q}_K\,, \tag{3.17}$$

with the numerical diffusion term

$$\boldsymbol{q}_K = \gamma\, \frac{\Lambda}{g}\, \max\{\Delta x, \Delta y\}\left(\omega \boldsymbol{u}_K^\perp + \nabla_K^e(f_e^K\, \phi_K)\right). \tag{3.18}$$

The numerical diffusion term based on the velocity divergence is the same as for the nine points scheme, see (3.10).

Semi-discrete energy : As for the first 9 points scheme, we are able to establish a discrete counterpart of (3.6) for the 13 points scheme (3.16). We do not detail the computations since they are very similar to the ones for the 9 points scheme in Proposition 3.4.3 and even simpler since the Coriolis term is defined on the cells and then obviously vanishes when multiplied by $\boldsymbol{u}_K$.

Proposition 3.4.7

The 13- points scheme (3.16) *satisfies the semi-discrete energy inequality*

$$\frac{\mathrm{d}}{\mathrm{dt}}\left(\sum_{K\in\mathbb{T}} m_K E_K\right) \leq \sum_{K\in\mathbb{T}} m_K \left[-\boldsymbol{q}_K \cdot (\omega \boldsymbol{u}_K^{\perp} + \nabla_K^e(f_e^K\, \phi_K)) - \pi_v \,\mathrm{div}_v^K\, \boldsymbol{u}_K\right].$$

Linearized well-balanced property : The linearized version of the scheme (3.16) reads

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{dt}} h_K + h_0\ \mathrm{div}_K^e(f_e^K\, \boldsymbol{u}_K) - h_0\, \mathrm{div}_K^e(f_e^K\, \tilde{\boldsymbol{q}}_K) = 0, \\ \dfrac{\mathrm{d}}{\mathrm{dt}} \boldsymbol{u}_K + \nabla_K^e(f_e^K\, \phi_K) - \nabla_K^v\, \tilde{\pi}_v = -\omega(\boldsymbol{u}_K - \tilde{\boldsymbol{q}}_K)^{\perp}, \end{cases} \tag{3.19}$$

with the numerical corrections

$$\tilde{\boldsymbol{q}}_K = \gamma\, \frac{\Lambda}{g h_0} \max\{\Delta x, \Delta y\} \left(\omega \boldsymbol{u}_K^{\perp} + \nabla_K^e(f_e^K\, \phi_K\right) \quad \text{and} \quad \tilde{\pi}_v = \nu\Lambda\, \max\{\Delta x, \Delta y\}\, \mathrm{div}_v^K\, \boldsymbol{u}_K.$$

Proposition 3.4.8

The linearized scheme (3.19) *preserves the discrete geostrophic equilibrium* $\tilde{\boldsymbol{q}}_K = 0$.

Proof : When the geostrophic balance expressed on the cells holds, i.e when $\omega \boldsymbol{u}_K^{\perp} + \nabla_K^e(f_e^K(\phi_K)) = 0$, we get

$$\mathrm{div}_K^e(f_e^K\, \boldsymbol{u}_K) = \frac{1}{\omega}\, \mathrm{div}_K^e\left(f_e^K \left((\nabla_K^e(f_e^K\, \phi_K))^{\perp}\right)\right)$$

and

$$\mathrm{div}_v^K\, \boldsymbol{u}_K = \frac{1}{\omega}\, \mathrm{div}_v^K \left((\nabla_K^e(f_e^K\, \phi_K))^{\perp}\right),$$

obvious computations show that these two quantities vanish. Hence the geostrophic balance expressed on the cells is included in the discrete kernel of the system. ■

Remark 3.4.9

As it contains a reconstruction operator on the water depth, the discrete geostrophic equilibrium $\tilde{\boldsymbol{q}}_K = 0$ *also contains spurious checkerboard solution*

$$\forall (i,j), \quad \boldsymbol{u}_{i,j} = 0 \quad \textit{and} \quad \begin{cases} \phi_{2i,2j} = a, \\ \phi_{2i,2j+1} = b, \\ \phi_{2i+1,2j} = c, \\ \phi_{2i+1,2j+1} = d. \end{cases}$$

Since it is constructed on a zero velocity mode, this spurious solution does belong to the kernel of the nonlinear scheme (3.16) *but remains bounded in time. Numerically, spurious solutions appear in the test cases 3.5.1, 3.5.3 and 3.5.4.*

3.4.3 Vertex-based 25 points scheme

The previous 13 points scheme (3.16) ensures the preservation of a discrete geostrophic equilibrium and the decreasing of the semi-discrete energy. But its kernel contains spurious solutions and we do not succeed to write a conservative version of the discrete pressure term. To adress this latter issue, here, we define the discrete geostrophic equilibrium and the perturbation q on the vertices v and then the perturbation π on the cells K. It leads to a scheme with a 25 points stencil.

Non-conservative semi-discrete scheme

With the discrete operators introduced in Section 3.3, the semi-discrete scheme reads :

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{dt}} h_K + \mathrm{div}_K^v \boldsymbol{\mathcal{F}}_v = 0, \\ \dfrac{\mathrm{d}}{\mathrm{dt}} (h_K \boldsymbol{u}_K) + \mathbf{div}_K^{v,up}(\boldsymbol{u}_K \otimes \boldsymbol{\mathcal{F}}_v) + h_K \, f_K^v(\nabla_v^K \phi_K) - f_K^v(\nabla_v^K \pi_K) = -\omega \left(h_K \, f_K^v(f_v^K \boldsymbol{u}_K) - f_K^v \boldsymbol{q}_v \right)^\perp, \end{cases} \tag{3.20}$$

where the interface fluxes are defined at the level of the vertices v as :

$$\boldsymbol{\mathcal{F}}_v = f_v^K(h_K \boldsymbol{u}_K) - \boldsymbol{q}_v \,, \tag{3.21}$$

with the numerical diffusion term on the flow rate

$$\boldsymbol{q}_v = \gamma \frac{\Lambda}{g} \max\{\Delta x, \Delta y\} (\omega \, f_v^K(\boldsymbol{u}_K^\perp) + \nabla_v^K \phi_K), \tag{3.22}$$

while the numerical diffusion term on the hydrostatic pressure is defined at the vertices

$$\pi_K = \nu \Lambda \, \max\{\Delta x, \Delta y\} h_K \, f_K^v(\mathrm{div}_v^K \boldsymbol{u}_K). \tag{3.23}$$

Remark 3.4.10

The term $f_K^v(\nabla_v^K \pi_K)$ *enforces the 25 points stencil. All the other operators are limited to a 9 points stencil.*

Semi-discrete energy : We first show the scheme (3.20) ensures a discrete counterpart of (3.6) through semi-discrete mechanic energy estimates. To do so, and as for the 9 points scheme, we first need the two following lemmas, describing the evolution of potential and kinetic energies.

Lemma 3.4.11 (*Semi-discrete scheme for the potential energy*)

We set $\mathcal{P}_K = \frac{1}{2} g\,(h_K)^2$ *for* $K \in \mathbb{T}$. *Then :*

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathcal{P}_K + \phi_K \,\mathrm{div}_K^v(f_v^K(h_K \boldsymbol{u}_K)) - \phi_K \,\mathrm{div}_K^v\, q_v = 0. \tag{3.24}$$

Lemma 3.4.12 (*Semi-discrete scheme for the kinetic energy*)

We set $\mathcal{K}_K = \frac{1}{2} h_K \|\boldsymbol{u}_K\|^2$ *for* $K \in \mathbb{T}$. *Then :*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}\mathcal{K}_K + \frac{1}{2}\,\mathbf{div}_K^{v,up}\left(\|\boldsymbol{u}_K\|^2 \otimes \boldsymbol{\mathcal{F}_v}\right) + h_K \boldsymbol{u}_K \cdot f_K^v\left(\nabla_K^v\,\phi_K\right)& \\ \leq -\omega(h_K \boldsymbol{u}_K \cdot f_K^v(f_v^K(\boldsymbol{u}_K^\perp)) - \boldsymbol{u}_K \cdot f_K^v(\boldsymbol{q}_v^\perp)) + \boldsymbol{u}_K \cdot f_K^v(\nabla_v^K\,\pi_K).&\end{aligned} \tag{3.25}$$

Proof: Using i) of Lemma 3.3.1 and the first equation of (3.20) we write :

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}\left(\frac{1}{2}h_K\|\boldsymbol{u}_K\|^2\right) &= \boldsymbol{u}_K \cdot \frac{\mathrm{d}}{\mathrm{dt}}(h_K \boldsymbol{u}_K) - \frac{1}{2}\|\boldsymbol{u}_K\|^2 \frac{\mathrm{d}}{\mathrm{dt}}h_K \\ &= -h_K \boldsymbol{u}_K \cdot f_K^v(\nabla_v^K\,\phi_K) - \boldsymbol{u}_K \cdot \mathbf{div}_K^{v,up}\left(\boldsymbol{u}_K \otimes \boldsymbol{\mathcal{F}_v}\right) + \omega \boldsymbol{u}_K \cdot (f_K^v\,\boldsymbol{q}_v)^\perp \\ &\quad + \frac{1}{2}\|\boldsymbol{u}_K\|^2\,\mathrm{div}_K^v\,\boldsymbol{\mathcal{F}_v} + \boldsymbol{u}_K \cdot f_K^v(\nabla_v^K\,\pi_K) - \omega h_K \boldsymbol{u}_K\, f_K^v(f_v^K(\boldsymbol{u}_K^\perp)).\end{aligned}$$

After some basic computations, we get the relation

$$\begin{aligned}&\frac{1}{2}\|\boldsymbol{u}_K\|^2\,\mathrm{div}_K^v\,\boldsymbol{\mathcal{F}}_v - \boldsymbol{u}_K \;\cdot\; \mathbf{div}_K^{v,up}\left(\boldsymbol{u}_K \otimes \boldsymbol{\mathcal{F}_v}\right) \\ &\qquad = -\frac{1}{2}\,\mathbf{div}_K^{v,up}(\|\boldsymbol{u}_K\|^2 \otimes \boldsymbol{\mathcal{F}_v}) + \frac{1}{2m_K}\sum_{e \subset \partial K}\|\boldsymbol{u}_K - \boldsymbol{u}_{Ke}\|^2\left((f_e^v\,\boldsymbol{\mathcal{F}_v}) \cdot \boldsymbol{n}_{e,k}\right)^-\,.\end{aligned}$$

The second term of the right hand side being non-positive, we get the announced result . ■

Proposition 3.4.13 (*Decreasing of the semi-discrete energy*)

We define the total energy $E_K = \mathcal{P}_K + \mathcal{K}_K$. *Then we obtain a discrete counterpart of* (3.6)

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{dt}}\left(\sum_{K \in \mathbb{T}} E_K\right) \\ &\leq -\max\{\Delta x, \Delta y\}\left[\sum_{K \in \mathbb{T}}\left[\nu \Lambda h_K\left(f_K^v(\mathrm{div}_v^K\,\boldsymbol{u}_K)\right)^2\right] + \sum_{v \in \mathbb{V}}\left[\gamma\,\frac{\Lambda}{g}\|\omega\, f_v^K(\boldsymbol{u}_K^\perp) + \nabla_v^K\,\phi_K\|^2\right]\right].\end{aligned} \tag{3.26}$$

Proof: Gathering relations (3.24) and (3.25), we obtain the following estimate for the total energy $E_K = \mathcal{P}_K + \mathcal{K}_K$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}E_K + \frac{1}{2}\,\mathbf{div}_K^{v,up}\left(\|\boldsymbol{u}_K\|^2 \otimes \boldsymbol{\mathcal{F}_v}\right) + \phi_K\,\mathrm{div}_K^v(f_v^K(h_K \boldsymbol{u}_K)) + h_K \boldsymbol{u}_K \cdot f_K^v\left(\nabla_v^K\,\phi_k\right)& \\ \leq -\omega(h_K \boldsymbol{u}_K \cdot f_K^v(f_v^K(\boldsymbol{u}_K^\perp)) - \boldsymbol{u}_K \cdot f_K^v(\boldsymbol{q}_v^\perp)) + \phi_K\,\mathrm{div}_K^v\,\boldsymbol{q}_v + \boldsymbol{u}_K \cdot f_K^v(\nabla_v^K\,\pi_K).&\end{aligned}$$

By telescoping and using periodic boundary condition we get :

$$\sum_{K\in\mathbb{T}} \left(\mathbf{div}_K^{v,up}\left(\|\boldsymbol{u}_K\|^2 \otimes \boldsymbol{\mathcal{F}_v}\right)\right) = 0.$$

Using iii) of Lemma 3.3.1 and v) of Lemma 3.3.3, we get :

$$\sum_{K\in\mathbb{T}} \left[\phi_K \,\mathrm{div}_K^v(f_v^K(h_K\boldsymbol{u}_K)) + h_K\boldsymbol{u}_K \cdot f_K^v\left(\nabla_v^K\,\phi_k\right)\right] = 0.$$

Thanks to iii) of Lemma 3.3.2 we get :

$$\sum_{K\in\mathbb{T}} \left[h_K\boldsymbol{u}_K \cdot f_K^v(f_v^K(\boldsymbol{u}_K^\perp)\right] = 0.$$

Finally, ii) and iii) of Lemma 3.3.3 and ii) of Lemma 3.3.2 leads us to the following semi-discrete inequality :

$$\Delta x\Delta y \frac{\mathrm{d}}{\mathrm{dt}}\left(\sum_{K\in\mathbb{T}} E_K\right) \leq -\Delta x\Delta y\left[\sum_{K\in\mathbb{T}}\left[\pi_K\, f_K^v(\mathrm{div}_v^K\,\boldsymbol{u}_K)\right] + \sum_{v\in\mathbb{V}}\left[\boldsymbol{q}_v\cdot(\omega\, f_v^K(\boldsymbol{u}_K^\perp) + \nabla_v^K\,\phi_K)\right]\right].$$

With the choices (3.22) and (3.23), we finally obtain the result. ■

Linearized well-balanced property : The linearized version of the scheme (3.20) is the following :

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{dt}}h_K + h_0\ \mathrm{div}_K^v\left(f_v^K\,\boldsymbol{u}_K\right) - h_0\,\mathrm{div}_K^v\,\tilde{\boldsymbol{q}}_v = 0, \\ \dfrac{\mathrm{d}}{\mathrm{dt}}\boldsymbol{u}_K + \ f_K^v(\nabla_v^K\,\phi_K) - \ f_K^v(\nabla_v^K\,\tilde{\pi}_K) = -\omega\left(f_K^v(f_v^K\,\boldsymbol{u}_K) - f_K^v\,\tilde{\boldsymbol{q}}_v\right)^\perp\ , \end{cases} \tag{3.27}$$

where

$$\tilde{\boldsymbol{q}}_v = \gamma\,\frac{\Lambda}{gh_0}\max\{\Delta x,\Delta y\}\left(\omega\, f_v^K(\boldsymbol{u}_K^\perp) + \nabla_v^K\,\phi_K\right) \quad \text{and} \quad \tilde{\pi}_K = \nu\Lambda\,\max\{\Delta x,\Delta y\}\, f_K^v(\mathrm{div}_v^K\,\boldsymbol{u}_K).$$

Proposition 3.4.14

The linearized scheme (3.27) *preserves the discrete geostrophic equilibrium* $\tilde{\boldsymbol{q}}_v = 0$.

Proof: When the geostrophic equilibrium holds on vertices, i.e. when $\omega\, f_v^K(\boldsymbol{u}_K^\perp) + \nabla_v^K\,\phi_K = 0$, we get

$$f_K^v(\omega\, f_v^K(\boldsymbol{u}_K^\perp) + \nabla_v^K\,\phi_K) = 0 \quad \text{and} \quad \mathrm{div}_K^v\left(f_v^K(\boldsymbol{u}_K)\right) = \mathrm{div}_K^v\left(\frac{(\nabla_v^K\,\phi_K)^\perp}{\omega}\right).$$

Basic computations then yield $\mathrm{div}_K^v\left(f_v^K(\boldsymbol{u}_K)\right) = 0$. Finally using property iii) of Lemma (3.3.3), we obtain that $f_K^v(\mathrm{div}_v^K\,\boldsymbol{u}_k) = 0$. Hence the corrections $\tilde{\boldsymbol{q}}_v$ and $\tilde{\pi}_K$ equal zero when the geostrophic equilibrium holds on the vertices and the scheme (3.27) preserves this equilibrium. Note that the natural divergence operator div_v^K fails to ensure $\mathrm{div}_v^K\,\boldsymbol{u_K} = 0$ when the geostrophic equilibrium holds at the vertices. It explains the choice of the expression of π_K in (3.23). ■

Remark 3.4.15

As it contains a reconstruction operator on both the velocity and the water depth, the discrete geostrophic equilibrium $\tilde{\boldsymbol{q}}_v = 0$ also contains spurious solutions

$$\forall (i,j), \quad \begin{cases} \boldsymbol{u}_{2i,2j} = \mathbf{a} \\ \boldsymbol{u}_{2i+1,2j+1} = \mathbf{b}, \\ \boldsymbol{u}_{2i,2j+1} = \mathbf{c} \\ \boldsymbol{u}_{2i+1,2j} = -\mathbf{a} - \mathbf{b} - \mathbf{c}, \end{cases} \quad \text{and} \quad \begin{cases} \phi_{2i,2j} = \phi_{2i+1,2j+1} = a, \\ \phi_{2i,2j+1} = \phi_{2i+1,2j} = b, \end{cases}$$

Due to the nonlinear term on the velocity, the spurious mode on the velocity does not belong to the kernel of the nonlinear scheme (3.14) but the checkerboard mode on the water depth does. However, the following spurious mode on the velocity does belong to the kernel of the nonlinear scheme :

$$\forall (i,j), \quad \begin{cases} \boldsymbol{u}_{2i+1,j} = \boldsymbol{u}_0, \\ \boldsymbol{u}_{2i,j} = -\boldsymbol{u}_0, \end{cases}$$

Numerically, small oscillations appear in the test case 3.5.4.

Conservative semi-discrete scheme

Here we propose a conservative form of the scheme (3.20), i.e. a version of the scheme with a conservative discrete pressure term as it should be in a finite volume framework. With the discrete operators introduced in Section 3.3, the conservative semi-discrete scheme reads :

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{dt}} h_K + \mathrm{div}_K^v \boldsymbol{\mathcal{F}}_v = 0, \\ \dfrac{\mathrm{d}}{\mathrm{dt}} (h_K \boldsymbol{u}_K) + \mathbf{div}_K^{e,up}(\boldsymbol{u}_K \otimes f_e^v \boldsymbol{\mathcal{F}}_v) + f_K^v \left(f_v^K(h_K)\, \nabla_v^K(\phi_K) \right) - f_K^v(\nabla_v^K \pi_K) \\ \qquad\qquad = -\omega \left(f_K^v(f_v^K(h_K)\, f_v^K(\boldsymbol{u}_K)) - f_K^v\, \boldsymbol{q}_v \right)^\perp, \end{cases} \tag{3.28}$$

where the interface fluxes are defined at the level of the vertices v as :

$$\boldsymbol{\mathcal{F}}_v = f_v^K(h_K)\, f_v^K(\boldsymbol{u}_K) - \boldsymbol{q}_v\,, \tag{3.29}$$

and the numerical diffusion terms $\boldsymbol{q}_v$ and π_K are defined by (3.22) and (3.23). Note the change. on the pressure term has for consequence a change on the mass flux and on the Coriolis term in order to preserve the energy balance.

Proposition 3.4.16

The pressure term $f_K^v(f_v^K(h_K)\,\nabla_v^K\,\phi_K)$ *is in conservative form*

$$[f_K^v(f_v^K(h_K)\,\nabla_v^K\,\phi_K)]_{i,j} =$$
$$\frac{g}{2}\begin{pmatrix} \dfrac{\left(\dfrac{h_{i+1,j+1}+h_{i+1,j}}{2}\right)^2+\left(\dfrac{h_{i+1,j}+h_{i+1,j-1}}{2}\right)^2}{2} - \dfrac{\left(\dfrac{h_{i-1,j+1}+h_{i-1,j}}{2}\right)^2+\left(\dfrac{h_{i-1,j}+h_{i-1,j-1}}{2}\right)^2}{2} \\ \hline 2\Delta x \\ \dfrac{\left(\dfrac{h_{i+1,j+1}+h_{i,j+1}}{2}\right)^2+\left(\dfrac{h_{i+1,j-1}+h_{i+1,j-1}}{2}\right)^2}{2} - \dfrac{\left(\dfrac{h_{i+1,j-1}+h_{i,j-1}}{2}\right)^2+\left(\dfrac{h_{i,j-1}+h_{i-1,j-1}}{2}\right)^2}{2} \\ \hline 2\Delta y \end{pmatrix}.$$

Semi-discrete energy : We now show the scheme (3.28) ensures a discrete counterpart of (3.6). We do not detail the computations since they are similar to the previous ones.

Lemma 3.4.17 (*Semi-discrete scheme for the potential energy*)

We set $\mathcal{P}_K = \dfrac{1}{2}g\,(h_K)^2$ *for* $K \in \mathbb{T}$. *Then :*

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathcal{P}_K + \phi_K\,\mathrm{div}_K^v(f_v^K(h_K)\,f_v^K(\boldsymbol{u}_K)) - \phi_K\,\mathrm{div}_K^v\,q_v = 0. \tag{3.30}$$

Lemma 3.4.18 (*Semi-discrete scheme for the kinetic energy*)

We set $\mathcal{K}_K = \dfrac{1}{2}h_K\|\boldsymbol{u}_K\|^2$ *for* $K \in \mathbb{T}$. *Then :*

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathcal{K}_K + \frac{1}{2}\,\mathbf{div}_K^{e,up}\left(\|\boldsymbol{u}\|^2\,f_e^v(\boldsymbol{\mathcal{F}}_v)\right) + \boldsymbol{u}_K\cdot f_K^v\left(f_v^K(h_K)\,\nabla_K^v\,\phi_K\right)$$
$$\leq -\omega(\boldsymbol{u}_K\cdot f_K^v(f_v^K(h_K)\,f_v^K(\boldsymbol{u}_K^\perp)) - \boldsymbol{u}_K\cdot f_K^v(\boldsymbol{q}_v^\perp)).\, + \boldsymbol{u}_K\cdot f_K^v(\nabla_v^K\,\pi_K), \tag{3.31}$$

Proposition 3.4.19 (*Decreasing of the semi-discrete energy*)

We define the total energy $E_K = \mathcal{P}_K + \mathcal{K}_K$. *It satisfies the discrete energy inequality* (3.26).

Proof: Gathering relations (3.30) and (3.31), we obtain the following estimate for the total energy $E_K = \mathcal{P}_K + \mathcal{K}_K$:

$$\frac{\mathrm{d}}{\mathrm{dt}}E_K + \frac{1}{2}\,\mathbf{div}_K^{e,up}\left(\|\boldsymbol{u}\|^2\,f_e^v(\boldsymbol{\mathcal{F}})\right) + \phi_K\,\mathrm{div}_K^e(f_e^K(f_v^K(h_K)\,f_v^K(\boldsymbol{u}_K))) + \boldsymbol{u}_K\cdot f_K^v\left(f_v^K(h_K)\,\nabla_v^K\,\phi_k\right)$$
$$\leq -\omega(\boldsymbol{u}_K\cdot f_K^v(f_v^K(h_K)\,f_v^K(\boldsymbol{u}_K^\perp)) - \boldsymbol{u}_K\cdot f_K^v(\boldsymbol{q}_v^\perp)) + \phi_K\,\mathrm{div}_K^v\,\boldsymbol{q}_v\, + \boldsymbol{u}_K\cdot f_K^v(\nabla_v^K\,\pi_K).$$

We get the result using properties mentioned in Lemmas 3.3.1, 3.3.2 and 3.3.3 and the particular form of the numerical diffusion terms. ■

Linearized well-balanced property : The linearized version of the conservative scheme (3.28) is the same as the non-conservative scheme (3.20). Hence, we obtain the same results : when the discrete geostrophic equilibrium holds at the vertices, the corrections $\tilde{\boldsymbol{q}}_v$ and $\tilde{\pi}_K$ equal zero and the geostrophic balance expressed on vertices is included in the kernel of the numerical scheme.

3.4.4 Time discretisation

For the discretisation in time, fluxes are taken explicit. Nevertheless it is well known that a fully explicit discretisation of the Coriolis term leads in that case to unstable schemes, see [16]. Then we consider the following discretisation of the Coriolis term for all the presented schemes :

$$\frac{u_x^{n+1}-u_x^n}{\Delta t}=\omega(\theta_u u_y^n+(1-\theta_u)u_y^{n+1}),\qquad \frac{u_y^{n+1}-u_y^n}{\Delta t}=-\omega(\theta_v u_x^n+(1-\theta_v)u_x^{n+1}),$$

with $\theta_u+\theta_v\leq 1$. Here we choose $\theta_u=1$ and $\theta_v=0$ so that the system is solved explicitly.

The time step is chosen following [5] such that

$$\Delta t^n\leq \min\left\{\frac{2}{\omega},\frac{\min(\Delta x,\Delta y)}{\max(\|\boldsymbol{u}^n\|+\sqrt{gh^n})}\right\}.$$

3.5 Numerical results

In the following, we present different test cases to highlight the comportment of the edge-based entropic scheme (3.7), the edge-based well-balanced scheme (3.14), cell-based scheme (3.16) and vertex-based scheme (3.20), compared to a Godunov-type scheme HLLC and the edge-based entropic scheme (3.7). The conservative vertex-based scheme (3.28) give identical results as the non-conservative version (3.20), thus this scheme is omitted of this section. We study the water depth h and the velocity vector field $\boldsymbol{u}=(u,v)$, as well as the energy of the schemes.

The numerical experiments are performed with the gravitational constant $g=1$ and the angular velocity $\omega=1$. The mesh is defined by a $[101\times 101]$ Cartesian grid and the numerical diffusion coefficients $\gamma=\nu=0.5$.

3.5.1 River test case

This test case is the counterpart of the lake-at-rest for classic shallow water equations. It consist of the simulation of a flow in a stationary state through a channel, with wall-type boundary condition for $x=-0.5$ and $x=0.5$, and periodic boundary condition for $y=-0.5$ and $y=0.5$. The initial condition (see Figure 3.3) is as follow :

$$\boldsymbol{u}=\begin{pmatrix}0\\ \dfrac{\epsilon}{\omega}\end{pmatrix}\quad\text{and}\quad \nabla h=\begin{pmatrix}\dfrac{\epsilon}{g}\\ 0\end{pmatrix},$$

with $\epsilon=0.01$.
Hence, this initial condition is both a stationary solution of the nonlinear Shallow Water equations with Coriolis (3.1) and a geostrophic equilibrium $g\nabla h+\omega\boldsymbol{u}^\perp=0$. We aim at numerical schemes being able to preserve this solution.

In Figures 3.4a and 3.4b, the vertex-based and edge-based well-balanced schemes preserve the initial condition. Due to the wall-type boundary condition the cell-based scheme (3.16) and the edge-based entropic scheme (3.7) present some anomalies on both the water depth and the velocities.

For the cell-based scheme, interferences on h and $\boldsymbol{u}$ appear. One may note that these interferences also occur naturally through the water-column test case (see Section 3.5.4) and are then preserved.

These interference are in the scheme's kernel (see Remark 3.4.9) and thus appear as a limitation of this scheme. Yet, one can remark that the mean values of h and $\boldsymbol{u}$ equal the stationary solution.

For the edge-based entropic scheme, defaults on h and $\boldsymbol{u}$ also appear on the wall boundaries but are not spread to the rest of the domain. Apart from theses defaults, the scheme preserve the stationary solution.

Finally, the HLLC scheme do not preserve the initial condition and will tend to the lake-at-rest state.

In Figure 3.5 the energy fully discrete is roughly preserved by our schemes (small variations except for the vertex-based), whereas it is decreasing for HLLC scheme.

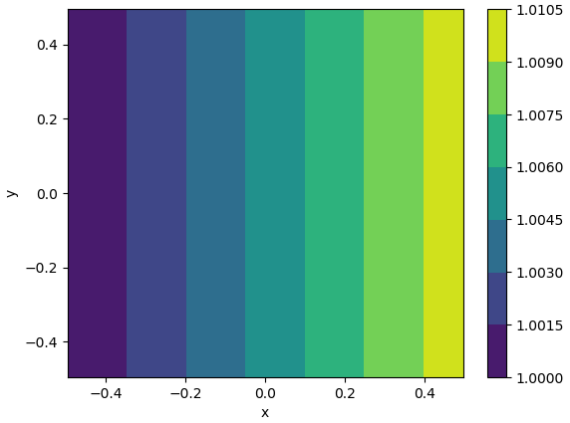
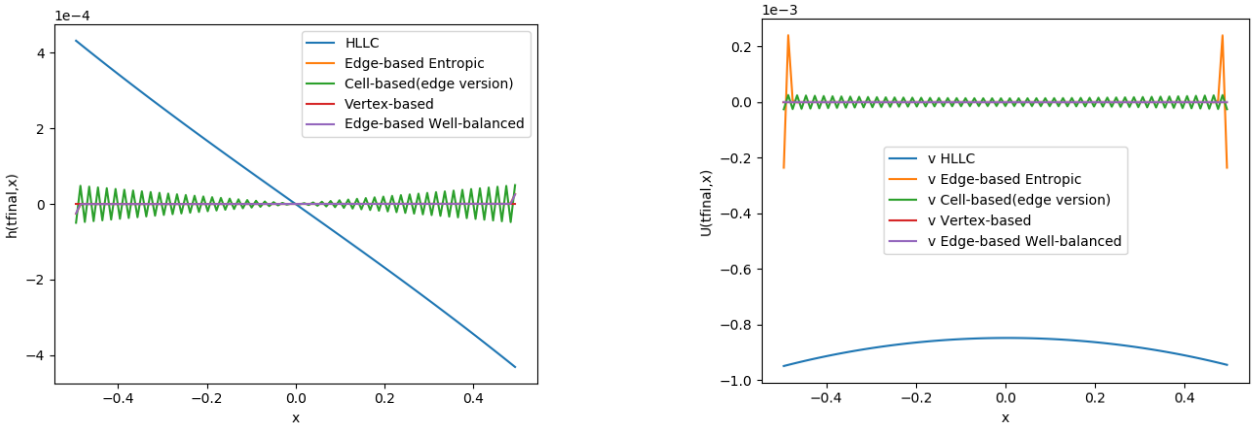


FIGURE 3.3 – Initial water depth.



(a) Difference on Water depth

(b) Difference on Velocity v

FIGURE 3.4 – Cross section in $y = 0$ of solution at $t = 20s$ difference to initial state.

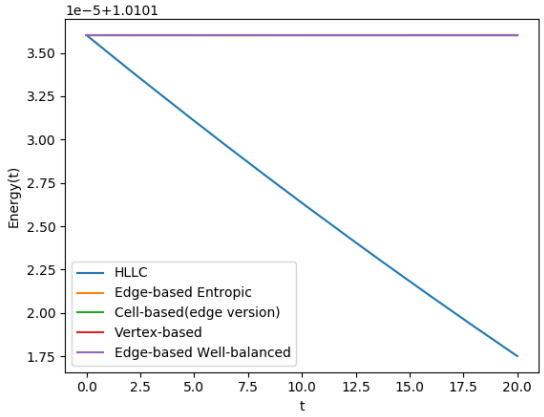


FIGURE 3.5 – Energy of the system in function of time.

3.5.2 Stationary vortex test case

We consider the test case introduced in [7] and defined by :

$$h(r) = \begin{cases} 1 + \dfrac{5\omega\epsilon}{2g} r^2 \text{ if } r \leqslant 0.2 \\ 1 + \dfrac{\omega\epsilon}{10g} - \dfrac{\omega\epsilon}{g}(0.3 - 2r + 2.5r^2) + \dfrac{\epsilon^2}{g}(3.5 - 20r + 12.5r^2 + 4\ln(5r)) \text{ if } 0.2 < r \leqslant 0.4 \\ 1 + \dfrac{\omega\epsilon}{5g} + \dfrac{\epsilon^2}{g}(4\ln(2) - 2.5) \text{ if } r > 0.4 \end{cases}$$

$$\text{and } \boldsymbol{u}(r,\theta) = \begin{cases} -5\epsilon\, r\, {}^t(\sin(\theta), \cos(\theta)) \text{ if } r \leqslant 0.2 \\ -(2 - 5r)\epsilon\, {}^t(\sin(\theta), \cos(\theta)) \text{ if } 0.2 < r \leqslant 0.4 \\ 0 \text{ if } r > 0.4 \end{cases}$$

where r and θ are the polar coordinates of a point of the domain. The parameter ϵ influences the initial water velocity of the vortex, which is linked to the gradient of the water depth through the geostrophic equilibrium. This create a stationary vortex that is a stationary solution of the nonlinear Shallow Water equations with Coriolis (3.1). Let us recall that this stationary solution is different from the geostrophic equilibrium (3.3) since the definition of the water depth contains a second term that is related to the nonlinear advective term in the equations. Nevertheless, when the coefficient ϵ introduced in the definition of the velocity field is small, the nonlinear term in the definition of the water depth is much smaller than the linear term and the stationary solution is very close to a geostrophic equilibrium. This initial condition can be seen on Figure 3.6. Finally, the domain has periodic boundary conditions. We take $\epsilon = 0.01$ and a final simulation time of $200s$. For this test case, the edge-based entropic, edge-based well-balanced, cell-based and vertex-based schemes all give similar results except the edge-based entropic one which slightly stand out. We thus choose to showcase only one scheme on the 2D graphs, the vertex-based one.

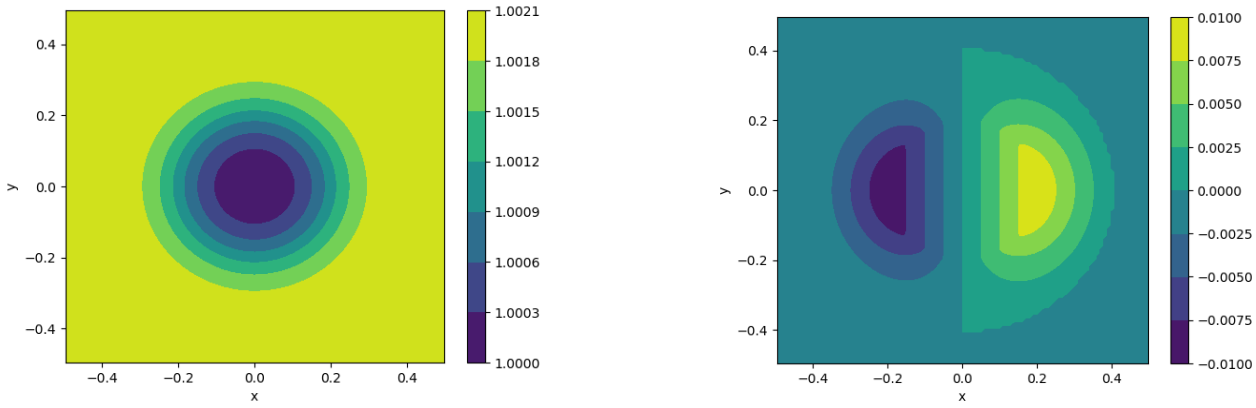
After $200s$, on Figures 3.7, 3.8 and 3.9 we can notice that the vertex-based scheme indeed mainly preserves the vortex. In that amount of time, the velocities and the depth of the vortex slightly reduced, but the overall shape of the vertex is preserved. Meanwhile, we can see that the HLLC scheme failed to preserve the said shape : the velocities and the depth reduced greatly, qualitatively, stratified velocity can be seen on Figure 3.8 whereas a distorted vortex appears on Figure 3.7 hence the scheme seems

to try to reach a lake at rest state.

In order to quantity the preservation of the stationary vortex, we introduce the following relative error :

$$\frac{|\min(h_{final}) - \min(h_{init})|}{\max(h_{init}) - \min(h_{init})},$$

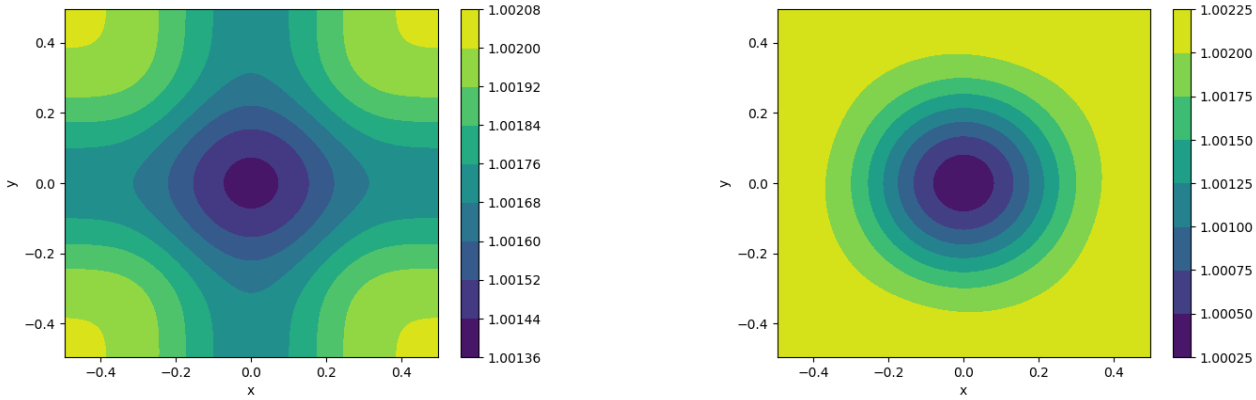
which track the bottom of the vortex. On Figure 3.10, we remark that the HLLC scheme quickly move away from the vortex, we can also note that the final water depth relative error made by the vertex-based scheme decreases with ϵ, while this parameter has no influence on the error for the HLLC scheme. Indeed, for any value of ϵ, the HLLC scheme has the same behaviour : the stationary vortex is not seen as an equilibrium and the difference of initial state thus has no effect while the scheme tries to reach the closest equilibrium in its kernel, the lake at rest. Meanwhile, the vertex-based scheme does consider the stationary vortex as an geostrophic equilibrium even if it is not a stationary solution for the scheme. Thus, when ϵ decrease the non-linear term in the scheme become negligible and the vortex appears as a quasi-equilibrium for the scheme.



(a) Water depth

(b) Velocity v

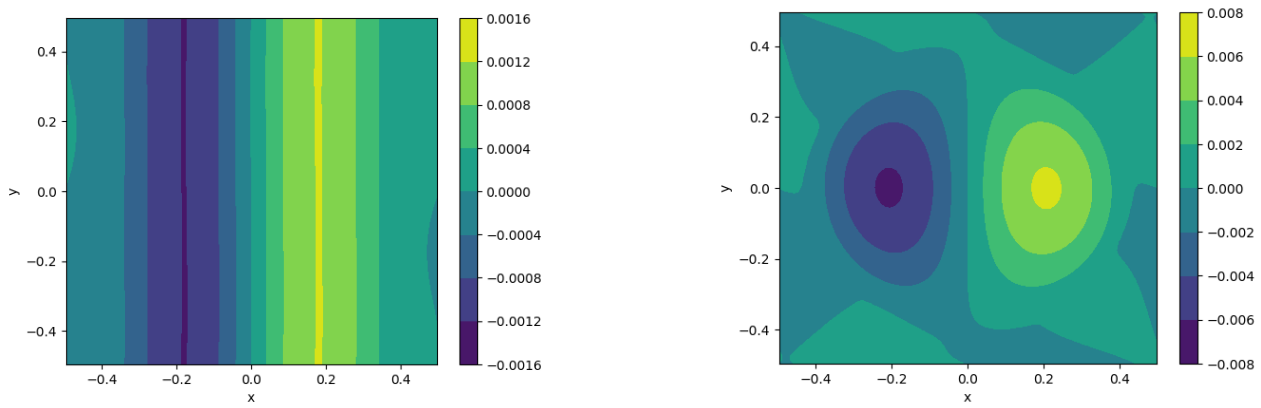
FIGURE 3.6 – Initial state of the stationary vortex.



(a) HLLC scheme

(b) Vertex-based scheme

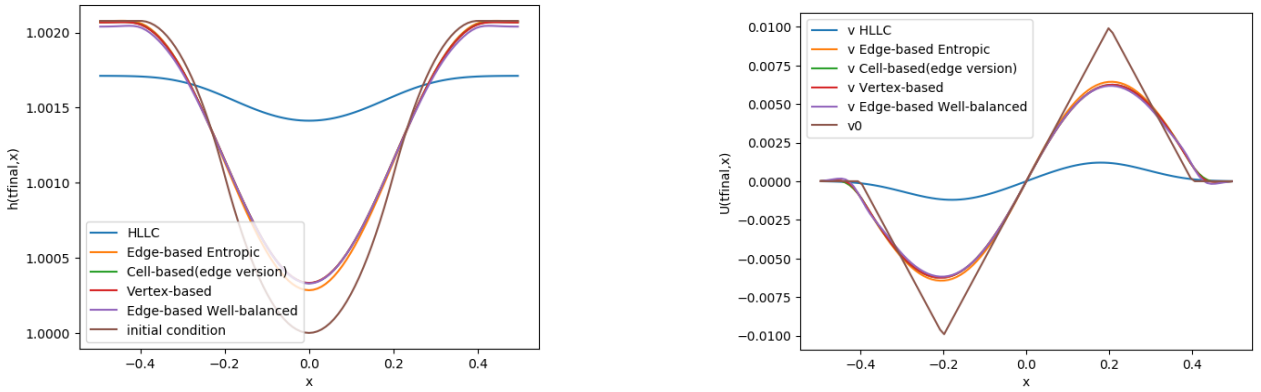
FIGURE 3.7 – Water depth at $t = 200s$.



(a) HLLC scheme

(b) Vertex-based scheme

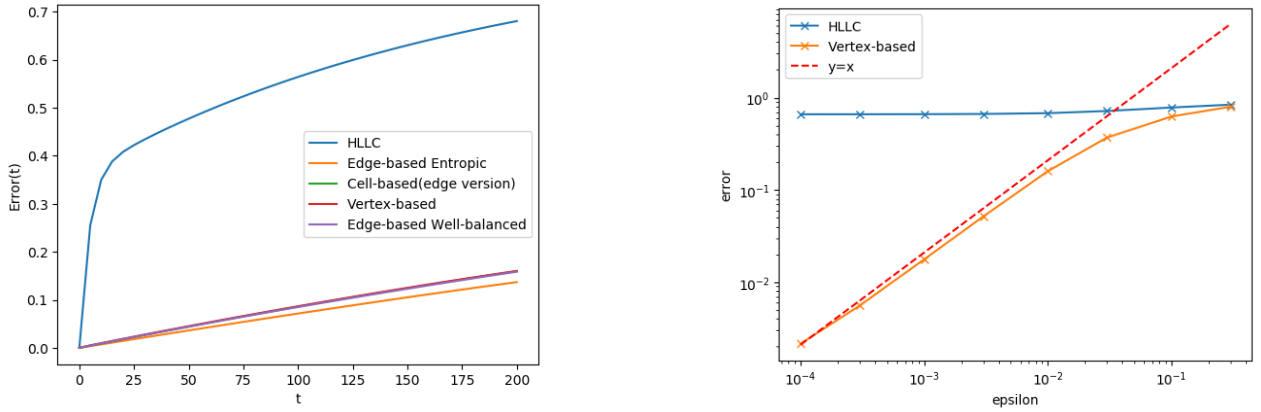
FIGURE 3.8 – Velocity v at $t = 200s$.



(a) Water depth

(b) Velocity v

FIGURE 3.9 – Cross sections $y = 0$ of solution at $t = 200s$.



(a) For the different scheme in function of time

(b) For the different ϵ at $t = 200s$

FIGURE 3.10 – Water depth relative error to initial state.

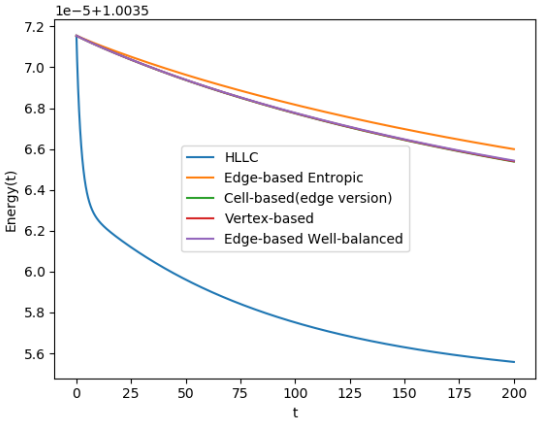


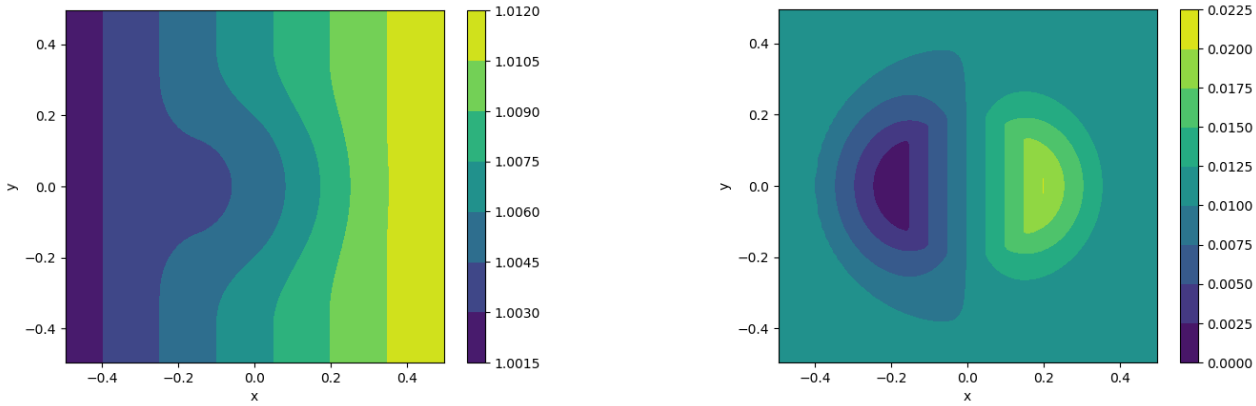
FIGURE 3.11 – Energy of the system in function of time for $\epsilon = 0.01$.

3.5.3 Translated vortex test case

This test case is the sum of two other ones : the river and the linear stationary vortex. Both are stationary solutions of the linear Shallow Water equations with Coriolis source term, and we are interested in the behaviour of our schemes when given that initial state (see Figure 3.12). As for the river test case, we enforce wall-type boundary condition for $x = -0.5$ and $x = 0.5$, and periodic boundary condition for $y = -0.5$ and $y = 0.5$. The slope of the river is $\epsilon = 0.01$, same as the vortex depth parameter.

After 20 seconds, we find in Figures 3.13 , 3.14 and 3.16 that the HLLC-scheme fails to preserve the initial state : the vortex is flattening out and the slope is decreasing. After 500s, it is brought back to the lake-at-rest. The other schemes also have a behaviour on par with the one they had on the two previous test cases and their graphs are overlapped. The vortex is preserved, except for a small depth reduction, and the slope of the river is constant, while interference appear for the cell-based scheme on Figure 3.16a. In addition, the river induces a translation of the vortex along the y axis.

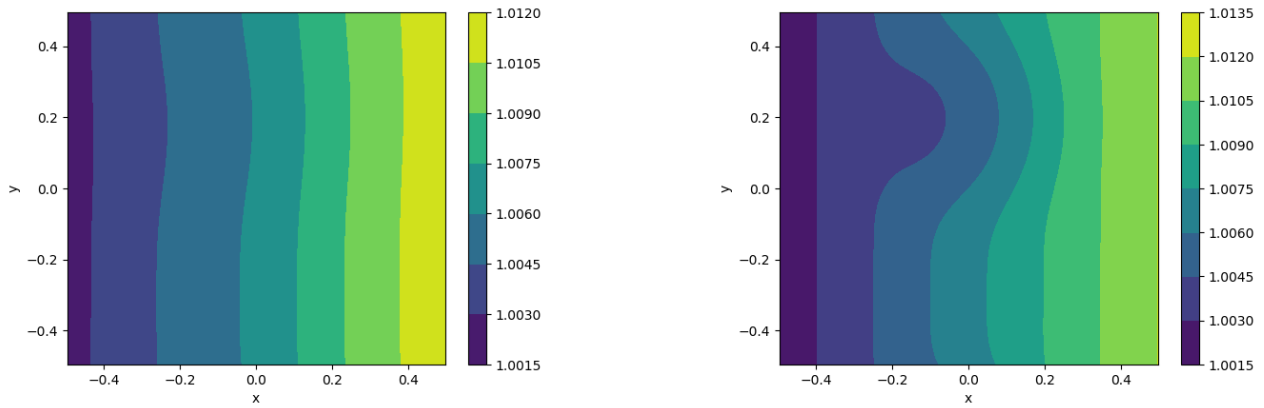
On Figure 3.18, we can note that the energy is also what one could expect from this superposition of solution : it is almost constant for the vertex-, cell- and edge-based entropic/well-balanced schemes, while it drops abruptly for the HLLC one.



(a) Water depth

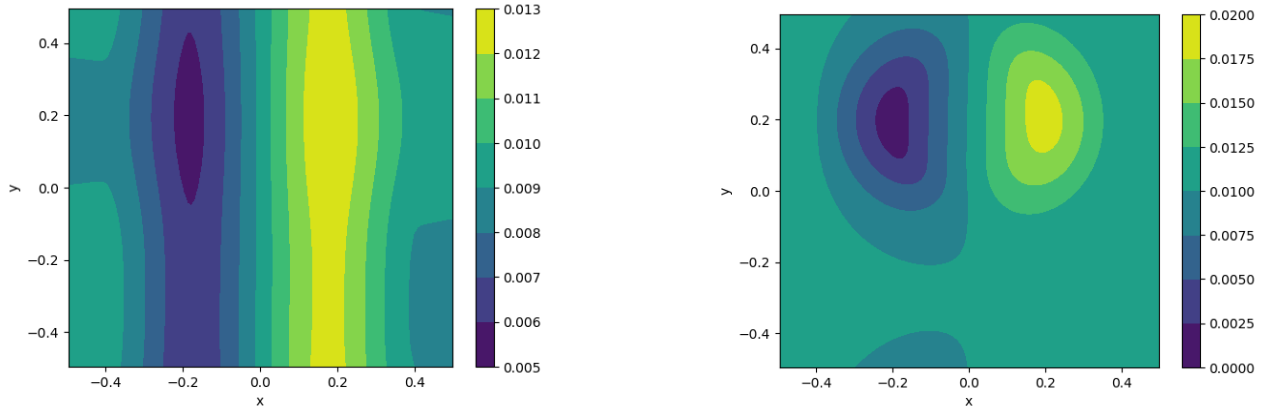
(b) Velocity v

FIGURE 3.12 – Initial state of the translated vortex.



(a) HLLC scheme (b) Vertex-based scheme

FIGURE 3.13 – Water depth at $t = 20s$.



(a) HLLC scheme (b) Vertex-based scheme

FIGURE 3.14 – Velocity v at $t = 20s$.

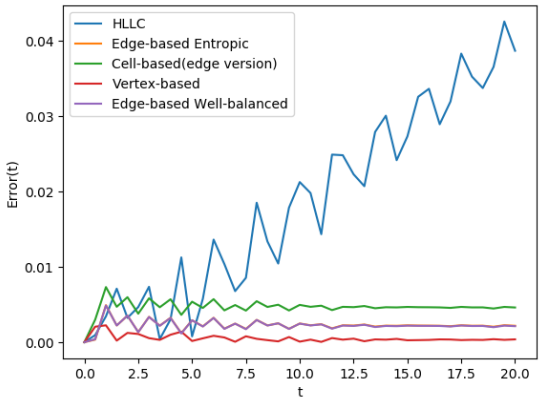
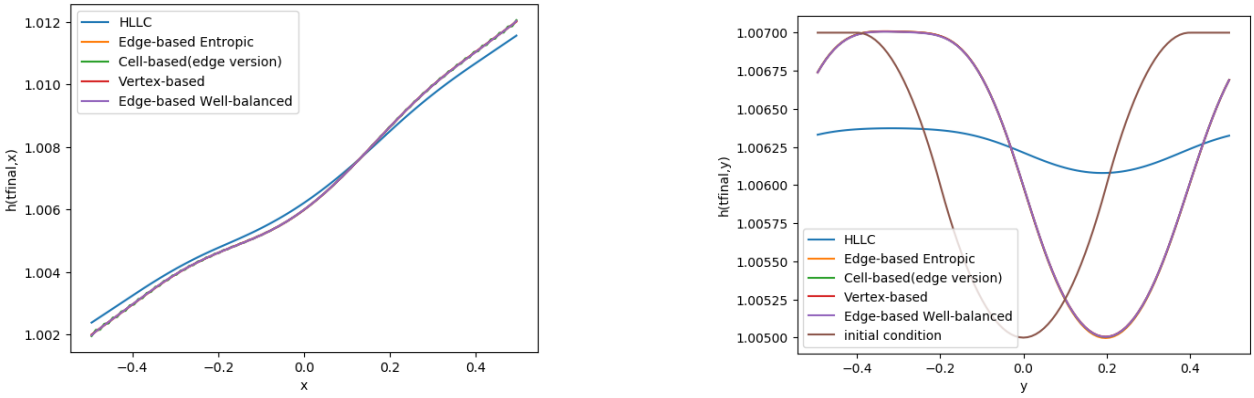
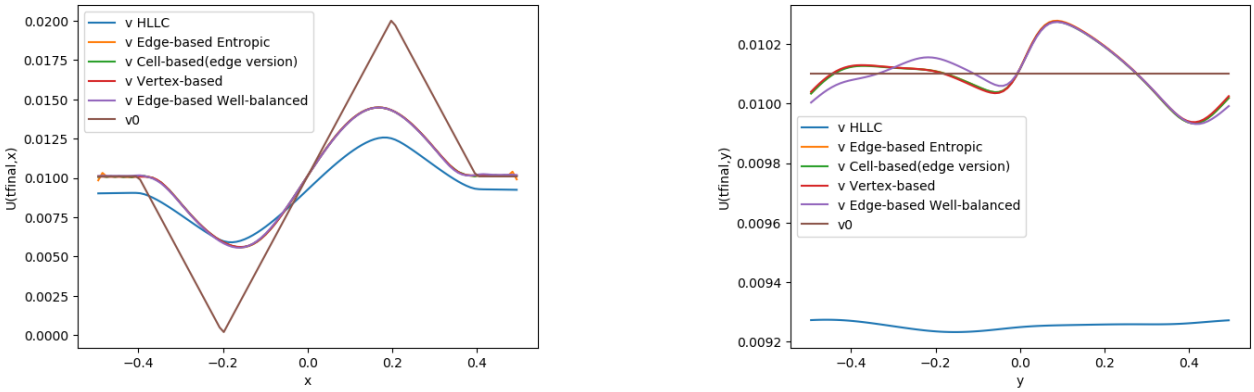


FIGURE 3.15 – Water depth relative error to initial state in function of time .



(a) $x = 0$ cross section (b) $y = 0$ cross section

FIGURE 3.16 – Cross section of water depth at $t = 20s$ in $x = 0$ (a) and $y = 0$ (b).



(a) $x = 0$ cross section (b) $y = 0$ cross section

FIGURE 3.17 – Cross section of velocity v at $t = 20s$ in $x = 0$ (a) and $y = 0$ (b).

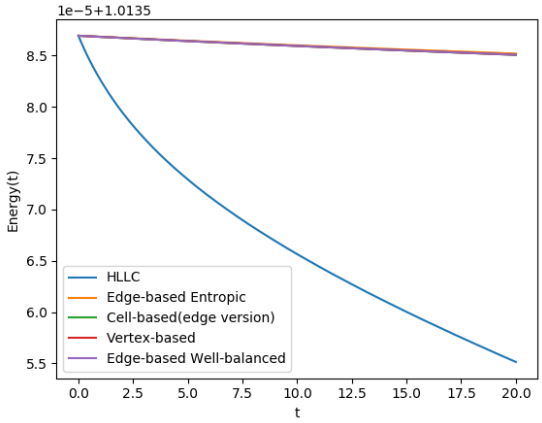


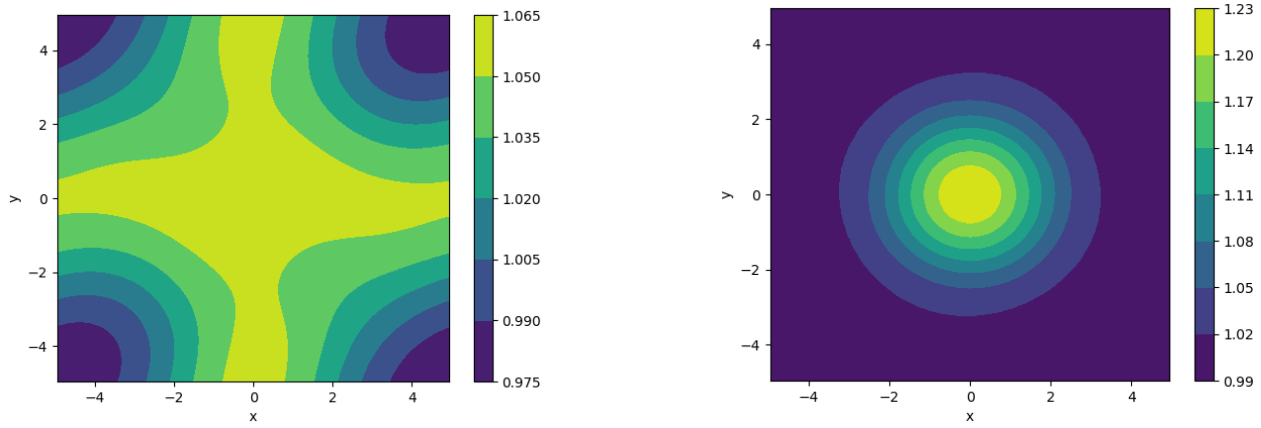
FIGURE 3.18 – Energy of the system in function of time for the different schemes.

3.5.4 Water-column test case

We consider a circular dam break with a radius of 1 and the domain has periodic boundary conditions. At $t = 0$, the velocity is zero throughout the domain and and the water height is 1 except for a water column in the center of height 2.

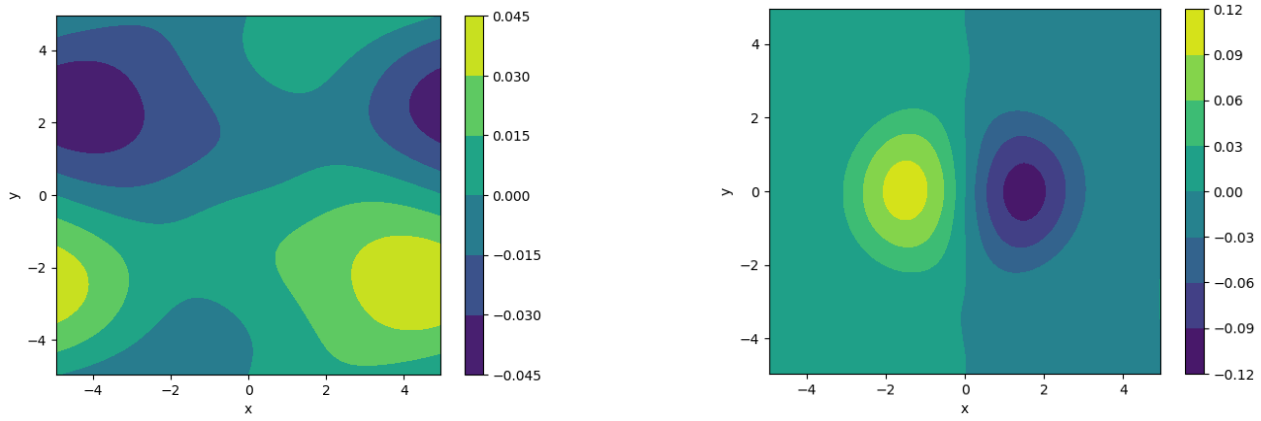
This initial condition is very far from a geostrophic equilibrium, thus on Figure 3.22 we can see that the energy drop is sharper than previously. The edge-based well-balanced, cell-based and vertex-based schemes have similar energy, their graphs are overlapped.

On Figure 3.21, all schemes have the same overall behaviour at the beginning. However, at time goes by, the HLLC-scheme gets to a lake at rest equilibrium, while the vertex-, cell- and edge-based entropic/well-balanced schemes stabilize around another geostrophic equilibrium. The graphs of the vertex-, cell- and edge-based well-balanced schemes are overlapped. At $t = 1s$, we can also note that interferences appear once again for the cell-based scheme, as well as for the vertex-based one. The edge-based entropic/well-balanced schemes seem immune to it for this test case.



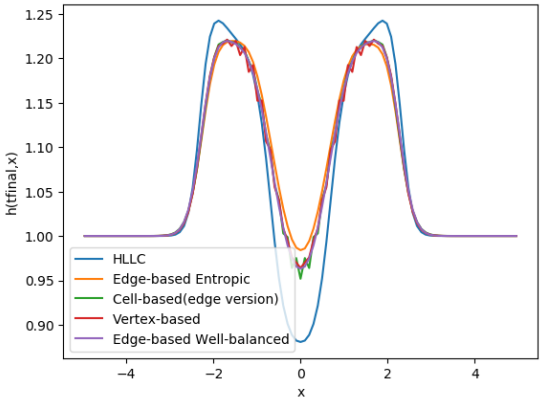
(a) HLLC scheme (b) Vertex-based scheme

FIGURE 3.19 – Water depth at $t = 100s$.

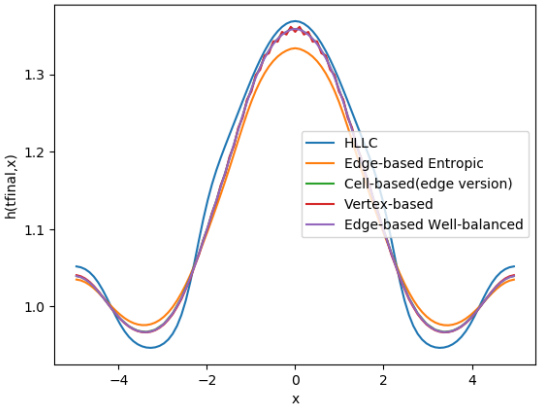


(a) HLLC scheme (b) Vertex-based scheme

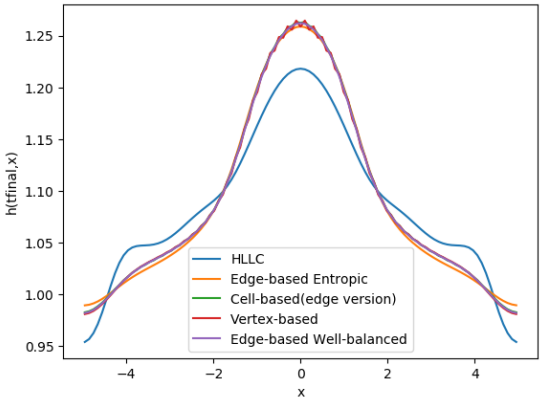
FIGURE 3.20 – Velocity v at $t = 100s$.



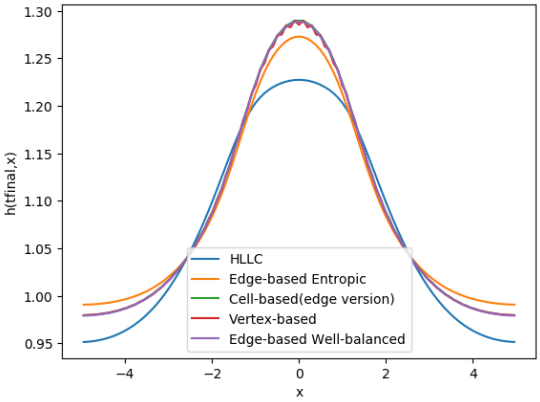
(a) $t = 1s$



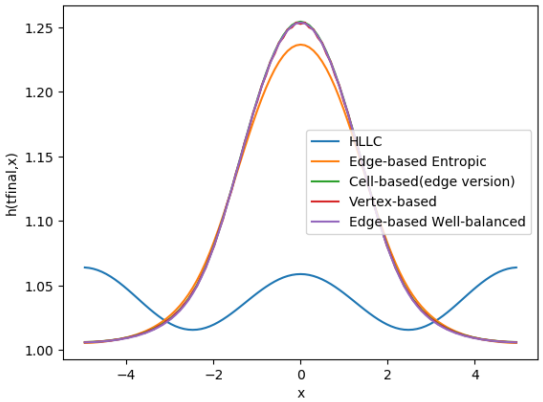
(b) $t = 4s$



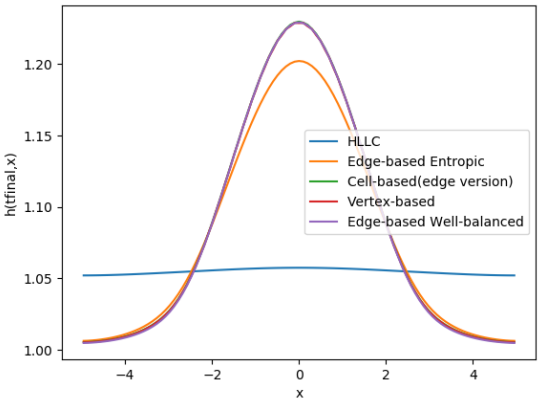
(c) $t = 6s$



(d) $t = 10s$



(e) $t = 40s$



(f) $t = 100s$

FIGURE 3.21 – Cross section in $y = 0$ of simulation results for the different schemes at different times.

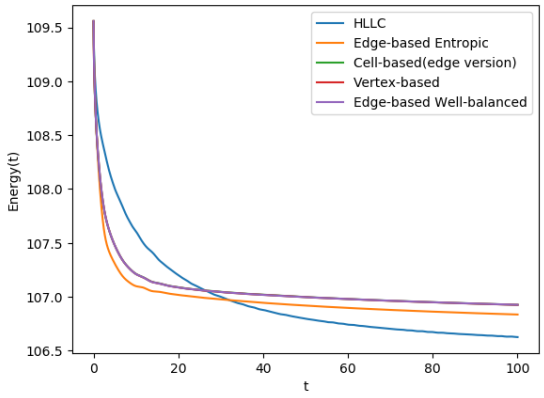


FIGURE 3.22 – Energy of the system in function of time for the different schemes.

3.6 Conclusion

In this work we have derived three semi-discrete schemes, namely edge- well-balanced, cell- and vertex-based for the shallow water equations with Coriolis force. For the edge-based well-balanced scheme we discretized the geostrophic equilibrium on the edges of the mesh and obtain a linearly well-balanced semi-discrete scheme. Although we failed to ensure the non-increasing energy, the numerical tests show a good short-term behaviour. For the cell-based scheme we discretized the geostrophic equilibrium on the cells of the mesh and obtain a linearly well-balanced semi-discrete scheme for which we have semi-discrete estimate. However, numerical results show that the kernel of this scheme contains spurious solutions and drive us to derived a final scheme. The vertex-based scheme has the same theoretical properties as the cell-based one and also has interference solutions but they appear less frequently.

Through different test cases, we saw that our schemes are more accurate around the geostrophic equilibrium than the HLLC one. However, spurious solutions appear for the cell- and vertex-based, further work has to be done in order to obtain a well-behaved scheme with restricted kernel. Also, the fully discrete energy of our schemes need to be studied.

Annexe A

Numerical approximation of the Shallow Water equations with Coriolis source term

E. Audusse[1], V. Dubos[2], A. Duran[3], N. Gaveau[4], Y. Nasseri[5], Y. Penel[2]

Abstract We investigate in this work a class of numerical schemes dedicated to the non-linear Shallow Water equations with topography and Coriolis force. The proposed algorithms rely on Finite Volume approximations formulated on collocated and staggered meshes, involving appropriate diffusion terms in the numerical fluxes, expressed as discrete versions of the linear geostrophic balance. It follows that, contrary to standard Finite-Volume approaches, the linear versions of the proposed schemes provide a relevant approximation of the geostrophic equilibrium. We also show that the resulting methods ensure semi-discrete energy estimates. Numerical experiments exhibit the efficiency of the approach in the presence of Coriolis force close to the geostrophic balance, especially at low Froude number regimes.

Contents

[1]LAGA, Institut Galilée, Université Sorbonne Paris Nord – 99 avenue Jean-Baptiste Clément, 93430 Villetaneuse.
[2]INRIA Paris - Sorbonne Université - CNRS (LJLL), team ANGE, 2 rue Simone Iff, CS 42112, 75589 Paris cedex 12.
[3]Institut Camille Jordan, Université Claude Bernard Lyon 1 – 43, boulevard du 11 novembre 1918, 69622 Villeurbanne.
[4]Institut Denis Poisson, Université D'Orléans, rue de Chartres, 45067 Orléans.
[5]Institut de Mathématiques de Marseille, Université d'Aix-Marseille – Technopôle Château Gombert, 13453, Marseille.

Introduction

The question of the accuracy of numerical schemes around stationary solutions or/and in asymptotic regimes has been a subject of great interest over the last two decades, see the seminal works [9,36,38] in late nineties and the books [13,34] ten years later. In the context of geophysical flows and more particularly for finite-volume methods applied to shallow water equations, a lot of works have been devoted to the accuracy around the *lake-at-rest* equilibrium and more recently extended to nonzero velocity one dimensional stationary states. The accuracy of numerical schemes around geostrophic equilibria which are of fundamental importance for large scales atmospheric or oceanographic flows was less investigated, see [53] for a general introduction to geophysical rotating fluid dynamics. To our knowledge, the first work in this field is due to Bouchut, Le Sommer and Zeitlin [14] but was fully accurate only for one-dimensional flows, as exhibited in [5] where accurate and stable Godunov type schemes were designed for linear two-dimensional shallow water equations. Recently two independent works [42,52] proposed an IMEX type scheme for fully nonlinear equations.

Here we aim at designing explicit schemes that are proved to be accurate and stable in the nonlinear framework. Our work is mostly based on the ideas developed in [5] for the linear rotating case and in [20,25] for the nonlinear non-rotating case. We investigate schemes applied to collocated and staggered schemes.

Let Ω be an open bounded domain of $\mathbb{R}^2$ and let $T > 0$. The nonlinear Shallow Water equations with Coriolis force formulated on $\Omega \times (0,T)$ read :

$$\begin{cases} \partial_t h + \mathrm{div}(h\boldsymbol{u}) = 0\,, \\ \partial_t(h\boldsymbol{u}) + \mathrm{div}(h\boldsymbol{u}\otimes\boldsymbol{u}) + h(\nabla\phi + \omega\boldsymbol{u}^\perp) = 0\,, \end{cases} \tag{A.1}$$

where h is the water height and $\boldsymbol{u} = (u_x, u_y)$ the horizontal velocity[1]. The Coriolis force is accounted for in the momentum equations through the angular speed ω. Following [20, 43], the pressure forces appear under a non conservative form through the scalar potential $\phi = gh$, where g is the standard gravity constant. For the sake of simplicity, a flat topography is considered in the present work, but the proposed approaches naturally extended to varying bottoms.[2]

It is well-known that the total energy of the system decomposes as $E = \mathcal{E} + \mathcal{K}$ where

$$\mathcal{E} = \frac{1}{2}gh^2 \quad \text{and} \quad \mathcal{K} = \frac{1}{2}h\|\boldsymbol{u}\|^2$$

stand respectively for potential and kinetic energies. We recall that the energy E plays the role of a mathematical entropy associated to the hyperbolic system (A.1) and regular solutions satisfy the following conservation law :

$$\partial_t E + \mathrm{div}\left[\left(\phi + \frac{1}{2}\|\boldsymbol{u}\|^2\right)h\boldsymbol{u}\right] = 0\,. \tag{A.2}$$

When developing numerical methods, main objectives are accuracy and stability. To get stability, a crucial objective is to build numerical approximations satisfying a discrete counterpart of (A.2) that ensures that the discrete energy is nonincreasing. To achieve this, a general strategy is to consider a sufficient amount of numerical diffusion in the scheme. But in some physical contexts such as low Froude number regimes or near specific stationary states, these diffusive terms may considerably degrade the accuracy of the approximations and specific methods are required. Here we are interested in flows around the geostrophic balance

$$\nabla\phi + \omega\boldsymbol{u}^\perp = 0, \quad \mathrm{div}\boldsymbol{u} = 0\,. \tag{A.3}$$

[1] $\boldsymbol{u}^\perp = (-u_y, u_x)$.

[2] In case of a non-flat topography z_b, $\phi = g(h + z_b)$.

To address such an issue, based on the linear case [5], and on the non-linear case without Coriolis force [20, 25], we propose a numerical approach involving discrete versions of these equilibria in the numerical fluxes. As a preliminary step, the strategy can be understood at the continuous level by investigating how the model (A.1) behaves with respect to some generic perturbations $(\boldsymbol{q}, \pi)$:

$$\begin{cases} \partial_t h + \mathrm{div}(h\boldsymbol{u} - \boldsymbol{q}) = 0\,, \\ \partial_t(h\boldsymbol{u}) + \mathrm{div}\big(\boldsymbol{u} \otimes (h\boldsymbol{u} - \boldsymbol{q})\big) + (h\nabla\phi - \nabla\pi) + \omega\,(h\boldsymbol{u} - \boldsymbol{q})^\perp = 0\,. \end{cases} \tag{A.4}$$

Hence $\boldsymbol{q}$ and π must be respectively seen as (small) perturbations with respect to the flow rate and to the hydrostatic pressure. The solutions to the modified equations (A.4) satisfy the following energy balance :

$$\partial_t E + \mathrm{div}\left[\left(\phi + \frac{1}{2}\|\boldsymbol{u}\|^2\right)\,(h\boldsymbol{u} - \boldsymbol{q}) - \pi\boldsymbol{u}\right] = -\boldsymbol{q}\cdot\left(\nabla\phi + \omega\boldsymbol{u}^\perp\right) - \pi\mathrm{div}\boldsymbol{u}\,, \tag{A.5}$$

which motivates a choice for $\boldsymbol{q}$ and π involving resp. the quantities $\nabla\phi + \omega\,\boldsymbol{u}^\perp$ and $\mathrm{div}\boldsymbol{u}$. Let us remark that these quantities govern the geostrophic equilibrium (A.3) associated to system (A.1) linearized around the steady state $(\tilde{h}, \tilde{\boldsymbol{u}}) = (h_0, 0)$ for a constant h_0 :

$$\begin{cases} \partial_t h = -h_0\,\mathrm{div}\boldsymbol{u}\,, \\ \partial_t \boldsymbol{u} = -(\nabla\phi + \omega\,\boldsymbol{u}^\perp)\,. \end{cases}$$

From a numerical point of view, diffusion terms are thus expected to have regularizing effects in the sense that they allow to recover a discrete counterpart of (A.5). Moreover, such terms are intended to vanish close to the geostrophic equilibrium, which must improve the quality of the approximations in this regime. Hence, the present approach can be seen as a non-linear extension of the linear case [5].

The paper is split into two parts : in the first section, we present the two geometric frameworks (collocated *vs.* staggered), we build some discrete operators with mimetic properties and we propose two well-balanced numerical schemes whose semi-discrete (in time) versions are proved to be stable. The second section is dedicated to the assessment of those schemes by means of the stationary vortex test case.

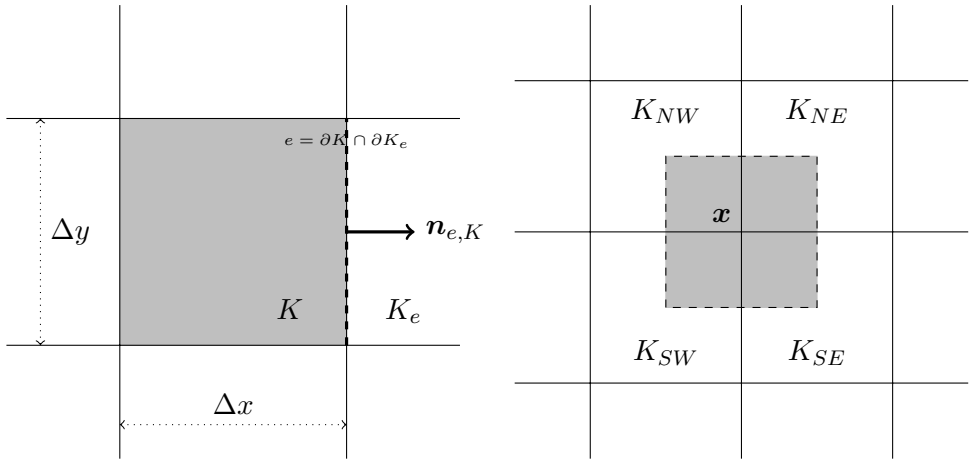
A.1 Design of the numerical schemes

A.1.1 Collocated scheme

Mesh and notations

To discretize System (A.1), we consider a tessellation $\mathbb{T}$ of the computational domain $\Omega \subset \mathbb{R}^2$ made of non-overlapping rectangular cells of sizes Δx, Δy. The set of all edges of the mesh is denoted by $\mathcal{E}$ and the set of vertices by $\mathcal{V}$.

- A generic cell of $\mathbb{T}$ is denoted by K, its area by m_K and its boundary by ∂K.
- The length of any edge e is denoted by m_e.
- Given a cell K and an edge $e \in \partial K$, K_e is the neighbouring cell to e (other than K) and $\boldsymbol{n}_{e,K}$ the outward normal pointing to K_e.
- Given a vertex $\boldsymbol{x}$, a dual cell is built by joining the centers of the neighbouring cells. Its area is denoted by $m_{\boldsymbol{x}}$.



(a) Primal cell K ; edge $e = \partial K \cap \partial K_e$; outward normal $\boldsymbol{n}_{e,K}$ to e pointing to K_e

(b) Dual cell associated to vertex $\boldsymbol{x}$

FIGURE A.1 – Geometric settings

Notations are pictured on Figs. A.1.

In particular, the discrete Green formula yields

$$\forall\, K \in \mathbb{T},\ \sum_{e \subset \partial K} m_e \boldsymbol{n}_{e,K} = \mathbf{0}. \tag{A.6}$$

In what follows, for any piecewise constant function φ we will denote

$$\overline{\varphi}_e = \frac{\varphi_{Ke} + \varphi_K}{2} \qquad \text{and} \qquad \overline{\varphi}_{\boldsymbol{x}} = \frac{1}{4}\left(\varphi_{K_{NW}} + \varphi_{K_{NE}} + \varphi_{K_{SE}} + \varphi_{K_{SW}}\right)$$

and introduce the vector quantity

$$\boldsymbol{\delta}\varphi_e = \frac{\varphi_{Ke} - \varphi_K}{2}\boldsymbol{n}_{e,K}.$$

In the same way, for a piecewise constant vector function $\boldsymbol{\Phi}$, we will note

$$\delta\boldsymbol{\Phi}_e = \frac{\boldsymbol{\Phi}_{Ke} - \boldsymbol{\Phi}_K}{2} \cdot \boldsymbol{n}_{e,K}.$$

Finally, the quantity

$$\varphi^{\pm} = \frac{1}{2}(\varphi \pm |\varphi|)$$

will refer to the the positive and negative parts of any scalar function φ. Before describing the schemes, we also need to define the following discrete divergence and gradient operators :

- Given a vector function $(\boldsymbol{\Phi}_e)_{e\in\mathcal{E}}$, the discrete divergence operator div_K reads

$$\forall\, K \in \mathbb{T},\ \mathrm{div}_K\, \boldsymbol{\Phi} = \frac{1}{m_K} \sum_{e \subset \partial K} m_e\, \boldsymbol{\Phi}_e \cdot \boldsymbol{n}_{e,K}. \tag{A.7}$$

  If $(\boldsymbol{\Phi}_K)_{K\in\mathbb{T}}$ is rather considered, then $\boldsymbol{\Phi}_e$ is replaced by $\overline{\boldsymbol{\Phi}}_e$ in (A.7) and the corresponding operator is denoted by $\mathrm{div}_K^c\, \boldsymbol{\Phi}$;

We can also define a divergence operator at the vertices when $(\mathbf{\Phi}_K)_{K\in\mathbb{T}}$ is considered by applying Formula (A.7) to the dual cell (see Fig.A.1(b)) :

$$\forall\, \boldsymbol{x}\in\mathcal{V},\ \mathrm{div}^c_{\boldsymbol{x}}\mathbf{\Phi} = \left(\frac{\mathbf{\Phi}_{K_{SE}}+\mathbf{\Phi}_{K_{NE}}}{2} - \frac{\mathbf{\Phi}_{K_{SW}}+\mathbf{\Phi}_{K_{NW}}}{2}\right)\cdot\frac{\boldsymbol{e}_1}{\Delta x} + \left(\frac{\mathbf{\Phi}_{K_{NE}}+\mathbf{\Phi}_{K_{NW}}}{2} - \frac{\mathbf{\Phi}_{K_{SE}}+\mathbf{\Phi}_{K_{SW}}}{2}\right)\cdot\frac{\boldsymbol{e}_2}{\Delta y}. \quad (A.8)$$

- In the same way, considering $(\mathbf{\Phi}_e)_{e\in\mathcal{E}}$ and $(\mathbf{\Psi}_K)_{K\in\mathbb{T}}$, the upwind divergence operator applied to the formal product $\mathbf{\Psi}\otimes\mathbf{\Phi}$ is defined by

$$\forall\, K\in\mathbb{T},\ \mathbf{div}^{up}_K(\mathbf{\Psi}\otimes\mathbf{\Phi}) = \frac{1}{m_K}\sum_{e\subset\partial K} m_e\left(\mathbf{\Psi}_K(\mathbf{\Phi}_e\cdot\boldsymbol{n}_{e,K})^+ + \mathbf{\Psi}_{K_e}(\mathbf{\Phi}_e\cdot\boldsymbol{n}_{e,K})^-\right). \quad (A.9)$$

  Likewise, if $(\mathbf{\Phi}_K)_{K\in\mathbb{T}}$ is rather considered, then $\mathbf{\Phi}_e$ is replaced by $\overline{\mathbf{\Phi}}_e$ in (A.9) and the corresponding operator is denoted by $\mathbf{div}^{c,up}_K(\mathbf{\Psi\Phi})$;

- Given a scalar function $(\phi_e)_{e\in\mathcal{E}}$, the discrete gradient reads

$$\forall\, K\in\mathbb{T},\ \nabla_K\phi = \frac{1}{m_K}\sum_{e\subset\partial K} m_e\,\phi_e\,\boldsymbol{n}_{e,K}. \quad (A.10)$$

  Likewise, if $(\phi_K)_{K\in\mathbb{T}}$ is rather considered, then ϕ_e is replaced by $\overline{\phi}_e$ in (A.10) and the corresponding operator is denoted by $\nabla^c_K\phi$.

  If $(\phi_{\boldsymbol{x}})_{\boldsymbol{x}\in\mathcal{V}}$ is considered, then ϕ_e is replaced by $\frac{1}{2}\sum_{\boldsymbol{x}\in\partial e}\phi_{\boldsymbol{x}}$ and the corresponding operator is denoted by $\nabla^x_K\phi$.

In particular, we have the mimetic properties (see [5])

$$\mathrm{div}^c_K(\nabla^\perp_K\varphi) = 0 \qquad \text{and} \qquad \sum_{K\in\mathbb{T}} m_K\nabla^x_K\varphi\cdot\mathbf{\Phi}_K = -\sum_{\boldsymbol{x}\in\mathcal{V}} m_{\boldsymbol{x}}\varphi_{\boldsymbol{x}}\mathrm{div}^c_{\boldsymbol{x}}\mathbf{\Phi}\ . \quad (A.11)$$

Numerical scheme

In this collocated version, all unknowns are located at the center of the cells except the perturbations ($\boldsymbol{q}$ located at the middle of the interfaces and π at the vertices). With the discrete operators (A.7-A.8-A.9-A.10) introduced above, the scheme reads, in the semi-discrete case :

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{dt}}h_K + \mathrm{div}_K(\boldsymbol{\mathcal{F}}) = 0, \\ \dfrac{\mathrm{d}}{\mathrm{dt}}(h_K\boldsymbol{u}_K) + \mathbf{div}^{up}_K(\boldsymbol{u}\otimes\boldsymbol{\mathcal{F}}) + h_K\,\nabla^c_K\phi - \nabla^x_K\pi = -\omega\left(h_K\boldsymbol{u}_K - \dfrac{1}{2}\displaystyle\sum_{e\subset\partial K}\boldsymbol{q}_e\right)^\perp, \end{cases} \quad (A.12)$$

where the interface fluxes are defined at the level of the edge e as :

$$\boldsymbol{\mathcal{F}}_e = \overline{h\boldsymbol{u}}_e - \boldsymbol{q}_e\,, \quad (A.13)$$

with

$$\boldsymbol{q}_e = \gamma\,\Delta t\,\overline{h}_e\ \left(2\frac{m_e}{m_K}\boldsymbol{\delta}\phi_e + \omega(\overline{\boldsymbol{u}}^\perp_e\cdot\boldsymbol{n}_{e,K})\boldsymbol{n}_{e,K}\right), \quad (A.14)$$

and

$$\pi_{\boldsymbol{x}} = \nu\Lambda\ \max\{\Delta x,\Delta y\}\overline{h}_{\boldsymbol{x}}\mathrm{div}^c_{\boldsymbol{x}}\boldsymbol{u}, \quad (A.15)$$

where γ and ν are positive dimensionless constants, and Λ is a positive characteristic velocity.[3]

[3] Typically, we take $\Lambda = \max_K\left\{\|\boldsymbol{u}^0_K\| + \sqrt{gh^0_K}\right\}$.

Remark A.1.1

Some remarks must be done on the formulation of the scheme :

1. *The origin of the expression* (A.14) *for the perturbation q will be explained in the proof below. We first mention that it is consistent with the geostrophic equilibrium (it is aimed to vanish at the equilibrium). Second, we only consider the normal component of $\overline{\boldsymbol{u}}_e$ in order to have the correct number of constraints. Morevoer, as the operator $\boldsymbol{\delta}\phi_e$ is along the normal vector, $\boldsymbol{q}_e = 0$ would imply (without the normal vector) a too restrictive kernel (characterised by $u_{K,x} = 0$ or $u_{K,y} = 0$ depending on the edges).*
2. *As $m_K = \Delta x\, \Delta y$ in this case, $\boldsymbol{q}_e$ does not actually depend on K.*
3. *Let us now explain why there is a $\frac{1}{2}$ coefficient the correction term in the right hand side of* (A.12)*. The previous remark combined to the cartesian structure induces that there are only two terms by component in the mean $\sum_e \boldsymbol{q}_e$.*[4]

Semi-discrete energy

We now show this scheme ensures a discrete counterpart of (A.5) through semi-discrete mechanic energy estimates. To obtain such a result, we first need the two following lemmas, describing the evolution of potential and kinetic energies.

Lemma A.1.2 (*Semi-discrete scheme for the potential energy*)

We set $\mathcal{E}_K = \dfrac{1}{2} g\,(h_K)^2$ for $K \in \mathbb{T}$. Then :

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathcal{E}_K + \mathrm{div}_K(\phi\boldsymbol{\mathcal{F}}) - h_K \boldsymbol{u}_K \cdot \nabla_K^c \phi + \mathcal{A}_K^{\mathcal{E}} = -\frac{1}{m_K}\sum_{e\subset\partial K} m_e\, \boldsymbol{\delta}\phi_e \cdot \boldsymbol{q}_e, \tag{A.16}$$

where the interface value for $\phi\boldsymbol{\mathcal{F}}$ is defined as :

$$(\phi\boldsymbol{\mathcal{F}})_e = \overline{\phi}_e \boldsymbol{\mathcal{F}}_e\,,$$

and the discrete conservative term :

$$\mathcal{A}_K^{\mathcal{E}} = \mathrm{div}_K\left(\boldsymbol{\delta}\phi\; \delta(h\boldsymbol{u})\right) = \frac{1}{m_K}\sum_{e\subset\partial K} m_e\, \boldsymbol{\delta}\phi_e\; \delta(h\boldsymbol{u})_e \cdot \boldsymbol{n}_{e,K}. \tag{A.17}$$

Proof: Using the definition of $\mathcal{E}_K$ and the mass conservation law (A.1), we get :

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathcal{E}_K = \phi_K \frac{\mathrm{d}}{\mathrm{dt}} h_K = -\phi_K\, \mathrm{div}_K(\boldsymbol{\mathcal{F}})\,.$$

Then, invoking the decomposition $\phi_K = \overline{\phi}_e + \dfrac{1}{2}\left(\phi_K - \phi_{K_e}\right)$ and the expression of the numerical fluxes (A.13) :

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathcal{E}_K = -\frac{1}{m_K}\sum_{e\subset\partial K} m_e\, \overline{\phi}_e \boldsymbol{\mathcal{F}}_e \cdot \boldsymbol{n}_{e,K} + \frac{1}{m_K}\sum_{e\subset\partial K} m_e\, \boldsymbol{\delta}\phi_e \cdot \overline{h\boldsymbol{u}}_e - \frac{1}{m_K}\sum_{e\subset\partial K} m_e\, \boldsymbol{\delta}\phi_e \cdot \boldsymbol{q}_e.$$

Lastly, using the relation $\overline{h\boldsymbol{u}}_e = h_K\boldsymbol{u}_K + \frac{1}{2}\left(h_{K_e}\boldsymbol{u}_{K_e} - h_K\boldsymbol{u}_K\right)$ and the discrete Green formula (A.6), basic computations yield :

$$\frac{1}{m_K}\sum_{e\subset\partial K} m_e\,\boldsymbol{\delta}\phi_e\cdot\overline{h\boldsymbol{u}}_e = h_K\boldsymbol{u}_K\cdot\nabla_K^c\phi + \frac{1}{m_K}\sum_{e\subset\partial K} m_e\,\boldsymbol{\delta}\phi_e\delta(h\boldsymbol{u})_e\cdot\boldsymbol{n}_{e,K},$$

which allows to conclude. ■

Lemma A.1.3 (*Semi-discrete scheme for the kinetic energy*)

We set $\mathcal{K}_K = \frac{1}{2}h_K\|\boldsymbol{u}_K\|^2$ *for* $K\in\mathbb{T}$. *Then :*

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathcal{K}_K + \frac{1}{2}\mathrm{div}_K^{up}\left(\|\boldsymbol{u}\|^2\boldsymbol{\mathcal{F}}\right) + h_K\boldsymbol{u}_K\cdot\nabla_K^c\phi - \mathcal{A}_K^{\mathcal{K}} \le -\frac{\omega}{2}\sum_{e\subset\partial K}\overline{\boldsymbol{u}}_e^{\perp}\cdot\boldsymbol{q}_e + \boldsymbol{u}_K\cdot\nabla_K^x\pi, \tag{A.18}$$

with anti-symmetric contributions given by :

$$\mathcal{A}_K^{\mathcal{K}} = \frac{\omega}{4}\sum_{e\subset\partial K}\left(\boldsymbol{u}_K - \boldsymbol{u}_{K_e}\right)\cdot\boldsymbol{q}_e^{\perp}\,. \tag{A.19}$$

Proof : We write :

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{dt}}\left(\frac{1}{2}h_K\|\boldsymbol{u}_K\|^2\right) = \boldsymbol{u}_K\cdot\frac{\mathrm{d}}{\mathrm{dt}}(h_K\boldsymbol{u}_K) - \frac{1}{2}\|\boldsymbol{u}_K\|^2\frac{\mathrm{d}}{\mathrm{dt}}h_K\\ &= -h_K\boldsymbol{u}_K\cdot\nabla_K^c\phi - \boldsymbol{u}_K\cdot\mathbf{div}_K^{up}\left(\boldsymbol{u}\otimes\boldsymbol{\mathcal{F}}\right) + \omega\boldsymbol{u}_K\cdot\left(\frac{1}{2}\sum_{e\subset\partial K}\boldsymbol{q}_e^{\perp}\right) + \frac{1}{2}\|\boldsymbol{u}_K\|^2\,\mathrm{div}_K(\boldsymbol{\mathcal{F}}) + \boldsymbol{u}_K\cdot\nabla_K^x\pi\,.\end{aligned} \tag{A.20}$$

First, we remark that :

$$\boldsymbol{u}_K\cdot\left(\frac{1}{2}\sum_{e\subset\partial K}\boldsymbol{q}_e^{\perp}\right) = \frac{1}{2}\sum_{e\subset\partial K}\overline{\boldsymbol{u}}_e\cdot\boldsymbol{q}_e^{\perp} + \frac{1}{4}\sum_{e\subset\partial K}\left(\boldsymbol{u}_K - \boldsymbol{u}_{K_e}\right)\cdot\boldsymbol{q}_e^{\perp} = -\frac{1}{2}\sum_{e\subset\partial K}\overline{\boldsymbol{u}}_e^{\perp}\cdot\boldsymbol{q}_e + \mathcal{A}_K^{\mathcal{K}}\,. \tag{A.21}$$

Then, after some basic computations, we get the relation

$$\frac{1}{2}\|\boldsymbol{u}_K\|^2\,\mathrm{div}_K(\boldsymbol{\mathcal{F}}) - \boldsymbol{u}_K\cdot\mathbf{div}_K^{up}\left(\boldsymbol{u}\otimes\boldsymbol{\mathcal{F}}\right) = -\frac{1}{2}\mathrm{div}_K^{up}(\|\boldsymbol{u}\|^2\boldsymbol{\mathcal{F}}) + \frac{1}{2m_K}\sum_{e\subset\partial K}\|\boldsymbol{u}_K - \boldsymbol{u}_{K_e}\|^2\left(\boldsymbol{\mathcal{F}}_e\cdot\boldsymbol{n}_{e,k}\right)^-\,. \tag{A.22}$$

The second term of the right hand side being non-positive, we get the announced result inserting (A.21) and (A.22) into (A.20). ■

Gathering relations (A.16) and (A.18), we obtain the following estimate for the total energy $E_K = \mathcal{E}_K + \mathcal{K}_K$:

$$\frac{\mathrm{d}}{\mathrm{dt}}E_K + \mathrm{div}_K(\phi\boldsymbol{\mathcal{F}}) + \frac{1}{2}\mathrm{div}_K^{up}(\|\boldsymbol{u}\|^2\boldsymbol{\mathcal{F}}) + \mathcal{A}_K^{\mathcal{E}} - \mathcal{A}_K^{\mathcal{K}} \le -\frac{1}{2}\sum_{e\subset\partial K}\left[2\frac{m_e}{m_K}\boldsymbol{\delta}\phi_e + \omega\overline{\boldsymbol{u}}_e^{\perp}\right]\cdot\boldsymbol{q}_e + \boldsymbol{u}_K\cdot\nabla_K^x\pi\,.$$

With the choices (A.14) and (A.15), we finally obtain a discrete counterpart of (A.5) thanks to (A.11).

Remark A.1.4

The contributions $\mathcal{A}_K^{\mathcal{E}}$ (A.17) and $\mathcal{A}_K^{\mathcal{K}}$ (A.19) involve anti-symmetric terms and therefore have no impact on the total energy budget :

$$\sum_{K\in\mathbb{T}} m_K(\mathcal{A}_K^{\mathcal{E}} - \mathcal{A}_K^{\mathcal{K}}) = 0,$$

up to boundary terms. In fact, these quantities can be interpreted as a bias to the conservative parts of potential and kinetic energies respectively.

Then :

Theorem A.1.5 (*Semi-discrete mechanic balance*)

The scheme (A.12, A.13, A.14, A.15) satisfies the following semi-discrete inequality :

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{dt}}\left(\sum_{K\in\mathbb{T}} m_K E_K\right)\\ &\leq -\frac{\gamma\Delta t}{2}\sum_{K\in\mathbb{T}} m_K \sum_{e\subset\partial K}\overline{h}_e\left\|\frac{\phi_{K_e}-\phi_K}{m_K/m_e} + \omega\overline{\boldsymbol{u}}_e^{\perp}\cdot\boldsymbol{n}_{e,K}\right\|^2 - \nu\Lambda\max\{\Delta x,\Delta y\}\sum_{\boldsymbol{x}\in\mathcal{V}} m_{\boldsymbol{x}}\overline{h}_{\boldsymbol{x}}(\mathrm{div}_{\boldsymbol{x}}^c\boldsymbol{u})^2\,.\end{aligned}$$

Steady states

Let us first assess the ability of the scheme to capture the lake at rest, characterized by $\phi = cte$ and $\boldsymbol{u} = 0$. One observes that the stabilization terms $\boldsymbol{q}_e$ and $\pi_{\boldsymbol{x}}$ vanish in such a configuration, leading directly to the exact preservation of the steady state in (A.12).

Then focus on the geostrophic equilibrium. In this case the nonlinear study is particularly complex and we choose to investigate the linear case. Although it leaves aside nonlinear interactions, this strategy may not seem irrelevant since most of low Froude number regimes encountered in practical contexts are governed by linear dynamics. Linearizing the scheme (A.12) around $\tilde{\boldsymbol{u}} = 0$ and $\tilde{h} = h_0$ for some constant $h_0 > 0$, we get :

$$\begin{cases}\dfrac{\mathrm{d}}{\mathrm{dt}}h_K + h_0\ \mathrm{div}_K^c\,\boldsymbol{u} - \mathrm{div}_K\,\boldsymbol{q} = 0,\\ \dfrac{\mathrm{d}}{\mathrm{dt}}\boldsymbol{u}_K + \nabla_K^c\phi - \nabla_K^{\boldsymbol{x}}\tilde{\pi} = -\omega\left(\boldsymbol{u}_K - \dfrac{1}{2}\displaystyle\sum_{e\subset\partial K}\boldsymbol{q}_e\right)^{\perp},\end{cases}$$

where

$$\boldsymbol{q}_e = \gamma\,\Delta t\,\left(2\frac{m_e}{m_K}\boldsymbol{\delta}\phi_e + \omega\,(\overline{\boldsymbol{u}}_e^{\perp}\cdot\boldsymbol{n}_{e,K})\boldsymbol{n}_{e,K}\right)\ \text{and}\ \tilde{\pi}_{\boldsymbol{x}} = -\nu\Lambda\max\{\Delta x,\Delta y\}\mathrm{div}_{\boldsymbol{x}}^c\boldsymbol{u}.$$

One should observe here that the discrete geostrophic equilibrium, characterized by :

$$\nabla_K^c\phi + \omega\boldsymbol{u}_K^{\perp} = 0 \overset{(A.11)}{\Longrightarrow} \mathrm{div}_K^c\,\boldsymbol{u}_K = 0$$

does not imply $\boldsymbol{q}_e = 0$ in the general case. Conversely, $\boldsymbol{q}_e = 0 \implies \pi_{\boldsymbol{x}} = 0$ according to [5] but does not imply $\nabla_K^c\phi + \omega\boldsymbol{u}_K^{\perp} = 0$. As exhibited in the numerical validations, although it does not exactly preserve the geostrophic balance, the present scheme behaves much better than standard Godunov-type approaches. However, as detailed in the next section, it is possible to correct such a failure by means of a staggered mesh.

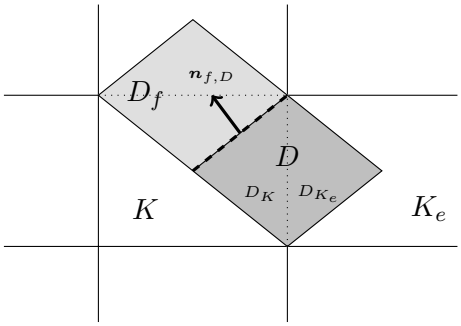


FIGURE A.2 – Geometric settings : for the primal mesh with cells K and K_e, refer to Figs. A.1 – for the dual mesh with diamond cells D and D_f separated by the interface $f = \partial D \cap \partial D_f$, $\boldsymbol{n}_{f,D}$ is the outward normal of f pointing to D_f.

A.1.2 Staggered scheme

Mesh and notations

On the basis of the primal mesh $\mathbb{T}$, a dual grid $\mathbb{T}^*$ can be built with respect to the Rannacher-Turek (RT) or Crouzeix-Raviart finite-element spaces, according to the formalism described in [3]. To set ideas, we focus here on the RT case considering a primal mesh $\mathbb{T}$ made of regular quadrilateral elements like in the previous section, but the following framework also applies to simplicial meshes or mixed triangular/quadrilateral elements.

- A dual cell $D \in \mathbb{T}^*$ is associated to each internal primal edge e as the quadrilateral admitting e as diagonal and the mass centers of the two neighbouring primal cells K and K_e as vertices (see Fig. A.2).[5]
- This dual element D is the union of the two triangles D_K, D_{K_e} separated by the edge e, whose areas will be denoted m_{D_K} and $m_{D_{K_e}}$, following the notations of the primal mesh.
- Given a dual cell D and an edge $e \subset \partial K$, D_f is the neighbouring cell to f (other than D) and $\boldsymbol{n}_{f,D}$ the outward normal pointing to D_f.

In this context, the water height evolves on the primal mesh $\mathbb{T}$ as for the previous scheme while the velocity field is located on the dual elements $D \in \mathbb{T}^*$. By analogy with the collocated frame, the upwind discrete divergence is defined on a dual element as :

$$\mathbf{div}_D^{up}(\boldsymbol{\Psi} \otimes \boldsymbol{\Phi}) = \frac{1}{m_D} \sum_{f \in \partial D} m_f \left(\boldsymbol{\Psi}_D (\boldsymbol{\Phi}_f \cdot \boldsymbol{n}_{f,D})^+ + \boldsymbol{\Psi}_{D_f} (\boldsymbol{\Phi}_f \cdot \boldsymbol{n}_{f,D})^- \right),$$

while the dual gradient is defined through the following formula :

$$\nabla_D \varphi = \frac{m_e}{m_D} (\varphi_{K_e} - \varphi_K) \, \boldsymbol{n}_{e,K} \, . \tag{A.23}$$

Note that this definition ensures the grad/div duality with respect to the L^2 inner product in the present staggered environment (for instance, see [37]) :

$$\sum_{K \in \mathbb{T}} m_K \varphi_K \, \mathrm{div}_K \, \boldsymbol{\Psi} = - \sum_{D \in \mathbb{T}^*} m_D \nabla_D \varphi \cdot \boldsymbol{\Psi}_D. \tag{A.24}$$

[5] The edge e and the corresponding diamond cell D will be used indifferently.

Numerical scheme

The scheme we consider is the following :

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{dt}} h_K + \mathrm{div}_K(\boldsymbol{\mathcal{F}}) = 0, \\ \dfrac{\mathrm{d}}{\mathrm{dt}} (h_D \boldsymbol{u}_D) + \mathbf{div}_D^{up}(\boldsymbol{u} \otimes \boldsymbol{\mathcal{F}}) + h_D \nabla_D \phi - \nabla_D \pi = -\omega \left(h_D \boldsymbol{u}_D - \boldsymbol{q}_D \right)^{\perp} , \end{cases} \tag{A.25}$$

where the primal fluxes are defined by interface by :

$$\boldsymbol{\mathcal{F}}_e = h_D \boldsymbol{u}_D - \boldsymbol{q}_D \,, \tag{A.26}$$

with

$$\boldsymbol{q}_D = \gamma \Delta t \; h_D \left(\nabla_D \phi + \omega \boldsymbol{u}_D^{\perp} \right) , \tag{A.27}$$

and

$$\boldsymbol{\pi}_K = \nu \Lambda \; \max\{\Delta x, \Delta y\} h_K \mathrm{div}_K \boldsymbol{u}. \tag{A.28}$$

As in the collocated frame, γ and ν are positive constants and Λ is consistent with a velocity. Considering the momentum equations (A.25) and the diffusion term (A.27), it is necessary to define a dual water height h_D. Following [3], we compute

$$m_D h_D = m_{D_K} h_K + m_{D_{K_e}} h_{K_e}.$$

It is then possible to choose dual fluxes $\boldsymbol{\mathcal{F}}_f$ (involved in the divergence term of the momentum equations (A.25)) so that a discrete law of mass conservation also holds on diamond cells :

$$\frac{\mathrm{d}}{\mathrm{dt}} h_D + \frac{1}{m_D} \sum_{f \in \partial D} m_f \boldsymbol{\mathcal{F}}_f \cdot \boldsymbol{n}_{f,D} = 0 \,,$$

which can be rewritten in the more compact form :

$$\frac{\mathrm{d}}{\mathrm{dt}} h_D + \mathrm{div}_D(\boldsymbol{\mathcal{F}}) = 0 \,. \tag{A.29}$$

As detailed in [3], these dual fluxes are expressed as linear combinations of the primal ones. These relations are purely geometrical and formulated locally, depending on the nature of the staggered mesh. Such a relation is mandatory to derive a kinetic energy budget as shown in the next paragraph.

Remark A.1.6

Let us notice that in (A.27)*, both components are taken into account in the definition of* $\boldsymbol{q}_D$ *unlike* (A.14) *where only the normal component was involved. In the present case,* $\boldsymbol{q}_D = 0$ *implies that the normal component of the velocity field vanishes (since the gradient is normal to the interface) but the value of interest is the reconstructed mean velocity*

$$\boldsymbol{u}_K = \frac{1}{4} \sum_{e \subset \partial K} \boldsymbol{u}_D.$$

Here $\boldsymbol{u}_D$ *is an unknown in itself (unlike in the collocated case where* $\overline{u}_e$ *is reconstructed from the unknowns* $\boldsymbol{u}_K$*).*

Semi-discrete energy

Indeed, defining the dual kinetic energy by $\mathcal{K}_D = \frac{1}{2} h_D \|\boldsymbol{u}_D\|^2$ and using the relation

$$\frac{\mathrm{d}}{\mathrm{dt}} \mathcal{K}_D = \boldsymbol{u}_D \cdot \frac{\mathrm{d}}{\mathrm{dt}}(h_D \boldsymbol{u}_D) - \frac{1}{2} \|\boldsymbol{u}\|^2 \frac{\mathrm{d}}{\mathrm{dt}} h_D$$

we have, with (A.29) and applying the relation (A.22) on the dual mesh :

$$\frac{\mathrm{d}}{\mathrm{dt}} \mathcal{K}_D + \mathrm{div}_D^{up} \left(\|\boldsymbol{u}_D\|^2 \boldsymbol{\mathcal{F}} \right) + h_D \boldsymbol{u}_D \cdot \nabla_D \phi \leq -\omega \boldsymbol{u}_D^{\perp} \cdot \boldsymbol{q}_D + \boldsymbol{u}_D \cdot \nabla_D \pi \,. \tag{A.30}$$

Then, looking at the potential energy, basic computations give :

$$\frac{\mathrm{d}}{\mathrm{dt}} \mathcal{E}_K + \mathrm{div}_K(\phi \boldsymbol{\mathcal{F}}) - \frac{1}{m_K} \sum_{e \subset \partial K} m_e \boldsymbol{\delta} \phi_e \cdot h_D \boldsymbol{u}_D = -\frac{1}{m_K} \sum_{e \subset \partial K} m_e \boldsymbol{\delta} \phi_e \cdot \boldsymbol{q}_D \,. \tag{A.31}$$

We are now left with the estimates (A.30) and (A.31), which are not expressed on the same mesh. A convenient way to understand the global impact of these quantities on the energy budget is to integrate them over their respective meshes, setting :

$$E = \sum_{K \in \mathbb{T}} m_K \mathcal{E}_K + \sum_{D \in \mathbb{T}^*} m_D \mathcal{K}_D \,,$$

and focus on the residual contributions at the level of dual elements. We first remark that, thanks to the duality relation (A.24), the second term of the right hand side of (A.30) furnish a first stabilizing term :

$$\sum_{D \in \mathbb{T}^*} m_D \boldsymbol{u}_D \cdot \nabla_D \pi = -\sum_{K \in \mathbb{T}} m_K \pi_K \, \mathrm{div}_K \, \boldsymbol{u},$$

according to the definition of π_K (A.28). Then, considering a generic element $D \in \mathbb{T}^*$, the non conservative terms in the kinetic energy balance (A.30) result in the following quantities on the left and right hand sides :

$$m_D h_D \boldsymbol{u}_D \cdot \nabla_D \phi \quad (LHS) \quad , \quad -m_D \omega \boldsymbol{u}_D^{\perp} \cdot \boldsymbol{q}_D \,. \quad (RHS)$$

On the other hand, the term issuing from (A.31) needs to be counted twice (contributions of the elements K and K_e sharing the edge e), which yields :

$$-2 m_e \boldsymbol{\delta} \phi_e \cdot h_D \boldsymbol{u}_D \quad (LHS) \quad , \quad -2 m_e \boldsymbol{\delta} \phi_e \cdot \boldsymbol{q}_D \,. \quad (RHS)$$

Using the definition (A.23), we easily verify that the terms of the left hand side are in exact balance and the remaining term reads :

$$-m_D \boldsymbol{q}_D \cdot \left(\nabla_D \phi + \omega \boldsymbol{u}_D^{\perp} \right) .$$

Replacing $\boldsymbol{q}_D$ by its expression (A.27), we have thus established the following result :

Theorem A.1.7 (*Semi-discrete mechanic balance*)

The scheme (A.25-A.26-A.27) satisfies the following semi-discrete inequality :

$$\frac{\mathrm{d}}{\mathrm{dt}} E \leq -\gamma \Delta t \sum_{D \in \mathbb{T}^*} m_D h_D \left\| \nabla_D \phi + \omega \boldsymbol{u}_D^{\perp} \right\|^2 - \nu \Lambda \, \max\{\Delta x, \Delta y\} \sum_{K \in \mathbb{T}} m_K h_K \left(\mathrm{div}_K \boldsymbol{u} \right)^2 .$$

As a remark, defining the quantity $\mathcal{K}_K = \frac{1}{m_K} \sum_{e \subset \partial K} m_D \mathcal{K}_D$, which can be understood as a kinetic energy associated to the cell K, the previous result can be reformulated under local energy estimates. Note that the present formalism embeds the implicit and explicit approaches recently proposed in [25] without Coriolis force.

Steady states

The linearized version of the scheme (A.25) is the following :

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{dt}} h_K + h_0 \ \mathrm{div}_K \boldsymbol{u} - \mathrm{div}_K \boldsymbol{q} = 0, \\ \dfrac{\mathrm{d}}{\mathrm{dt}} \boldsymbol{u}_D + \nabla_D \phi - \nabla_D \tilde{\pi} = -\omega \left(\boldsymbol{u}_D - \boldsymbol{q}_D\right)^{\perp} , \end{cases}$$

where

$$\boldsymbol{q}_D = \gamma \, \Delta t \left(\nabla_D \phi + \omega \, \boldsymbol{u}_D^{\perp} \right) \quad \text{and} \quad \tilde{\pi}_K = \nu \Lambda \, \max\{\Delta x, \Delta y\} \mathrm{div}_K \boldsymbol{u}.$$

From this, we immediately see that discrete kernel of the system corresponds to the geostrophic balance expressed on the dual grid $\mathbb{T}^*$:

$$\nabla_D \phi + \omega \, \boldsymbol{u}_D^{\perp} = 0 \, ,$$

and is consequently exactly preserved by the linearized scheme. We may indeed check that $\mathrm{div}_K(\nabla_D^{\perp} \phi) = 0$ by construction which implies $\mathrm{div}_K \boldsymbol{u} = 0$ when the geostrophic equilibrium holds.

A.1.3 Time discretisation

For the discretisation in time, fluxes are taken explicit. We only remind the reader of the standard discretisation of the right hand side (Coriolis force). The ODE

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \omega v, \qquad \frac{\mathrm{d}v}{\mathrm{d}t} = -\omega u,$$

is solved by means of the θ scheme

$$\frac{u^{n+1} - u^n}{\Delta t} = \omega(\theta_u v^n + (1-\theta_u) v^{n+1}), \qquad \frac{v^{n+1} - v^n}{\Delta t} = -\omega(\theta_v u^n + (1-\theta_v) u^{n+1})$$

with $\theta_u + \theta_v \leq 1$. Here we choose $\theta_u = 1$ and $\theta_v = 0$ so that the system is solved explicitly.

The time step is chosen following [5] such that

$$\Delta t^n \leq \min \left\{ \frac{2}{\omega}, \frac{\min(\Delta x, \Delta y)}{\max_D \|\boldsymbol{u}_D^n\| + \sqrt{g h_D^n}} \right\}.$$

A.2 Numerical assessments of the schemes

A.2.1 Presentation of the test cases

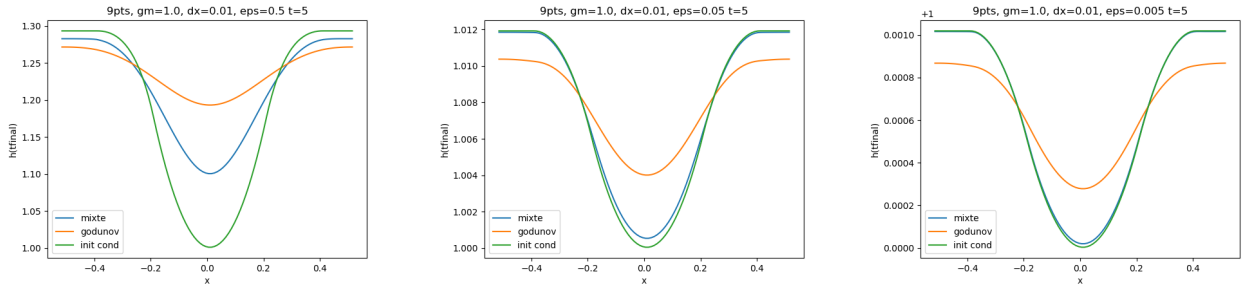
To assess the numerical schemes designed in this work, we consider the stationary vortex described in [7]. Set in the domain $\Omega = [-0.5, 0.5] \times [-0.5, 0.5]$, the equations are supplemented with the initial velocity field given in polar coordinates

$$\boldsymbol{u}^0(r, \theta) = \nu_\theta(r) \boldsymbol{e}_\theta \quad \text{where} \quad \nu_\theta(r) = \varepsilon \left[5r \mathbf{1}_{\{r<1/5\}} + (2 - 5r) \mathbf{1}_{\{1/5 \leq r < 2/5\}} \right].$$

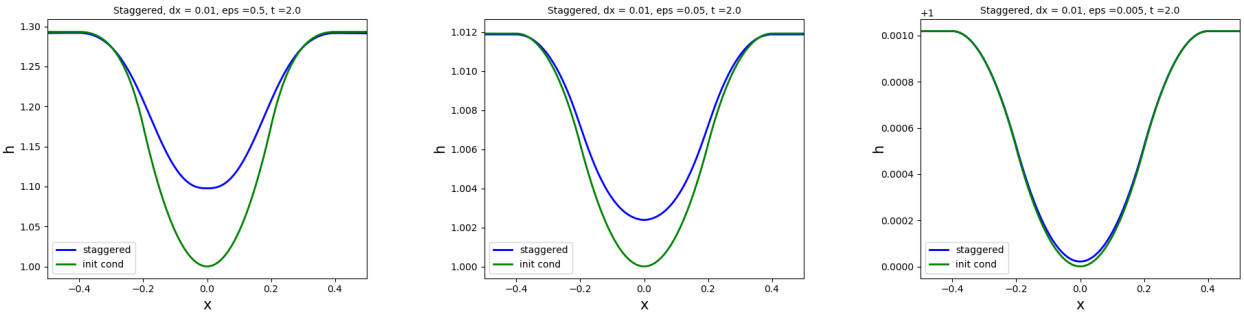
The initial water height is then computed so that the steady version of the nonlinear equations is satisfied, *i.e.* such that

$$\partial_r h^0(r, \theta) = \frac{\nu_\theta(r)}{g} \left(\omega + \frac{\nu_\theta(r)}{r} \right).$$

As mentioned in [5], this case induces a Froude number of order $\mathcal{O}(\varepsilon)$.



(a) $\varepsilon = 0.5$ (b) $\varepsilon = 0.05$ (c) $\varepsilon = 0.005$



(d) $\varepsilon = 0.5$ (e) $\varepsilon = 0.05$ (f) $\varepsilon = 0.005$

FIGURE A.3 – Stationary vortex : cross-section of the water height at final time for different values of ε ; comparisons between the exact solution (*green curves*), the numerical solution obtained with the classic Godunov scheme (*orange curve*), the collocated scheme (A.12) (A-B-C) and the staggered scheme (A.25) (D-E-F) (*blue curves*)

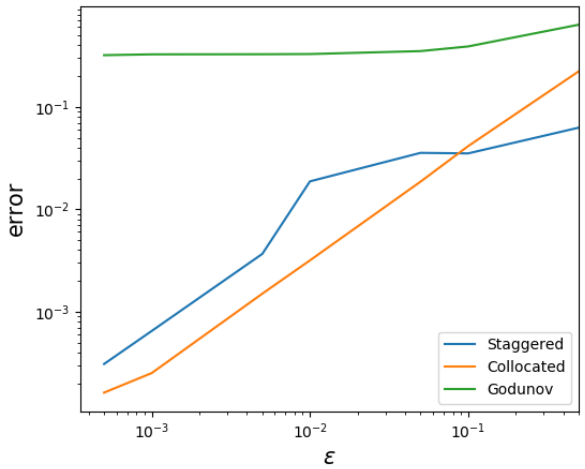


FIGURE A.4 – Error in L^2 norm between the numerical solution and the exact solution at final time

A.2.2 Numerical results

Given the initial condition prescribed in Section A.2.1, we compare the robustness of several numerical schemes. The data are a steady solution of the nonlinear equations but their approximation on the grid is not necessarily preserved by any numerical scheme. We remind that the present paper was aimed at designing a scheme that is robust in the low Froude number regime by means of diffusion terms based on linear equilibria.

We observe on Figure A.3 that the smaller ε, the more accurate the solution obtained with the collocated scheme (A.12) while the accuracy of the standard Godunov scheme is barely improved.

This is confirmed by Fig. A.4 where a kind of relative error is plotted :

$$\mathrm{E}(\varepsilon) = \frac{\sum_i (h_K^N - h_K^0)^2}{\sum_i (h_K^0 - \overline{h})^2}$$

where N is the index of the final time and $\overline{h} = \max_K h_K^0$. The Godunov scheme broadly provides results of which the accuracy is independent from ε while the modifications proposed in this paper leads to a second order scheme.

These results tend to show that the target is reached insofar as the resulting scheme is robust in the low Froude number regime where the standard Godunov scheme was inaccurate. Considering the presented simulations, it does not seem that the two constants γ and ν must necessarily be both strictly positive to get stability. A linear stability analysis and other numerical investigations are currently in progress to clarify this point.

A.3 Conclusion

We managed to construct some corrections to the standard Godunov scheme so that the resulting collocated and staggered schemes are accurate and energy-decreasing around the geostrophic equilibrium in the low Froude number regime. The stationary vortex test case go to show the improvement due to these new terms. The proofs are given in the semi-discrete case. In particular, the semi-discrete energy inequality requires $\gamma \geq 0$ and $\nu \geq 0$ so that it decreases. Future works will thus deal with the stability analysis of the fully discrete version. Moreover, further numerical investigations are needed to ensure the schemes behave well in a large range of applications.

Acknowledgements. The authors express their deep gratitude to the French National Research Agency project NABUCO, grant ANR-17-CE40-0025 and the french INSU-CNRS (Institut National des Sciences de l'Univers – Centre National de la Recherche Scientifique) program LEFE-MANU (Méthodes Mathématiques et Numériques), project DWAVE for the funding of the project, as well as to the organizing committee of the CEMRACS'19 event. The authors also thank Martin Parisot for fruitful discussions.

Bibliographie

[1] N. Aïssiouene. *Numerical analysis and discrete approximation of a dispersive shallow water model.* Theses, Université Pierre et Marie Curie - Paris VI, Dec. 2016.

[2] N. Aïssiouene, M.-O. Bristeau, E. Godlewski, A. Mangeney, C. Parés, and J. Sainte-Marie. A two-dimensional method for a family of dispersive shallow water models. *SMAI Journal of Computational Mathematics*, 6 :39, Sept. 2020.

[3] G. Ansanay-Alex, F. Babik, J.-C. Latché, and D. Vola. An L^2-stable approximation of the Navier–Stokes convection operator for low-order non-conforming finite elements. *International Journal for Numerical Methods in Fluids*, 66(5) :555–580, 2011.

[4] E. Audusse, F. Bouchut, M. Bristeau, R. Klein, and B. Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM Journal on Scientific Computing*, 25(6) :2050–2065, 2004.

[5] E. Audusse, M. Do, P. Omnes, and Y. Penel. Analysis of modified godunov type schemes for the two-dimensional linear wave equation with coriolis source term on cartesian meshes. *Journal of Computational Physics*, 373 :91–129, 2018.

[6] E. Audusse, V. Dubos, A. Duran, N. Gaveau, Y. Nasseri, and Y. Penel. Numerical approximation of the shallow water equations with Coriolis source term. *ESAIM : Proceedings*, 70 :31–44, June 2021.

[7] E. Audusse, R. Klein, and A. Owinoh. Conservative discretization of coriolis force in a finite volume framework. *Journal of Computational Physics*, 228(8) :2934–2950, 2009.

[8] J. Behrens and F. Dias. New computational methods in tsunami science. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 373(2053) :20140382, 2015.

[9] A. Bermudez and M. Vazquez-Cendon. Upwind Methods for Hyperbolic Conservation Laws with Source Terms. *Computers and Fluids*, 23(8) :1049–1071, 1994.

[10] C. Berthon, A. Duran, A., F. Foucher, K. Saleh, and J. De Dieu Zabsonré. Improvement of the hydrostatic reconstruction scheme to get fully discrete entropy inequalities. *Journal of Scientific Computing*, May 2019.

[11] C. Berthon, M. M'Baye, M. H. Le, and D. Seck. A well-defined moving steady states capturing Godunov-type scheme for Shallow-water model. *International Journal on Finite Volumes*, 15, 2020.

[12] P. Bonneton, E. Barthelemy, F. Chazel, R. Cienfuegos, D. Lannes, F. Marche, and M. Tissier. Recent advances in serre–green naghdi modelling for wave transformation, breaking and runup processes. *European Journal of Mechanics-B/Fluids*, 30(6) :589–597, 2011.

[13] F. Bouchut. *Nonlinear stability of finite volume methods for hyperbolic conservation laws, and well-balanced schemes for sources*, volume 2/2004. Birkhäuser Basel, 2004.

[14] F. Bouchut, J. Le Sommer, and V. Zeitlin. Frontal geostrophic adjustment and nonlinear wave phenomena in one-dimensional rotating shallow water. Part 2. High-resolution numerical simulations. *Journal of Fluid Mechanics*, 514 :35–63, 2004.

[15] M.-O. Bristeau, A. Mangeney, J. Sainte-Marie, and N. Seguin. An energy-consistent depth-averaged euler system : derivation and properties. *Discrete and Continuous Dynamical Systems - Series B*, 20(4) :28, 2015.

[16] M. Castro, J. Antonio López, and C. Pares. Finite volume simulation of the geostrophic adjustment in a rotating shallow-water system. *SIAM Journal on Scientific Computing*, 31 :444–477, 01 2008.

[17] F. Chazel, D. Lannes, and F. Marche. Numerical simulation of strongly nonlinear and dispersive waves using a green–naghdi model. *Journal of Scientific Computing*, 48(1) :105–116, 2011.

[18] A. Chertock, M. Dudzinski, A. Kurganov, and M. Lukácová-Medvid'ová. Well-balanced schemes for the shallow water equations with coriolis forces. *Numerische Mathematik*, 138 :939–973, 2018.

[19] A. Chorin. Numerical solution of the navier–stokes equations. *Mathematics of Computation*, 22, 10 1968.

[20] F. Couderc, A. Duran, and J.-P. Vila. An explicit asymptotic preserving low froude scheme for the multilayer shallow water model with density stratification. *Journal of Computational Physics*, 343 :235–270, 2017.

[21] M. W. Dingemans. *Water Wave Propagation Over Uneven Bottoms*. World Scientific Publishing Company, 1997.

[22] J. Droniou, R. Eymard, T. Gallouët, C. Guichard, and R. Herbin. *The gradient discretisation method*, volume 82 of *Mathématiques et Applications*. Springer International Publishing AG, Aug. 2018.

[23] J. Droniou, R. Eymard, T. Gallouët, and R. Herbin. Gradient schemes : a generic framework for the discretisation of linear, nonlinear and nonlocal elliptic and parabolic equations. *Mathematical Models and Methods in Applied Sciences*, pages 23(13), 2395–2432, 2013.

[24] J. Droniou, R. Eymard, T. Gallouët, and R. Herbin. A unified analysis of elliptic problems with various boundary conditions and their approximation. *Czechoslovak Mathematical Journal*, 70 :339–368, 2020.

[25] A. Duran, J.-P. Vila, and R. Baraille. Energy-stable staggered schemes for the shallow water equations. https ://hal.archives-ouvertes.fr/hal-01988382/, Jan. 2019.

[26] J. R. Edwards. Reflections on the early development of the "ausm family" of riemann solvers. *Shock Waves*, 29(5) :601–609, 2019.

[27] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*. Springer-Verlag New York, 2004.

[28] L. Euler. Principes généraux du mouvement des fluides. *Mémoires de l'académie des sciences de Berlin, Volume 11, pp. 274-315.*, 1757.

[29] R. Eymard, T. Gallouët, and R. Herbin. *Finite Volume Methods*, volume 7 of *Handbook of Numerical Analysis*. Elsevier, 2000.

[30] R. Eymard, T. Gallouët, and R. Herbin. RTk mixed finite elements for some nonlinear problems. . *Mathematics and Computers in Simulation*, pages 1–20, Dec. 2014.

[31] R. Eymard, C. Guichard, and R. Herbin. Small-stencil 3D schemes for diffusive flows in porous media. *ESAIM : Mathematical Modelling and Numerical Analysis*, 46 :265–290, 2012.

[32] V. Girault and P.-A. Raviart. *Finite Element Methods for Navier-Stokes Equations*. 1986.

[33] S. Glimsdal, G. K. Pedersen, C. B. Harbitz, and F. Lovholt. Dispersion of tsunamis : does it really matter ? *Natural Hazards and Earth System Sciences*, 13(6) :1507–1526, 2013.

[34] L. Gosse. *Computing qualitatively correct approximations of balance laws*, volume 2. Springer, 2013.

[35] A. E. Green and P. M. Naghdi. A derivation of equations for wave propagation in water of variable depth. *Journal of Fluid Mechanics*, 78(2) :237–246, 1976.

[36] J. Greenberg and A. LeRoux. A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM Journal on Numerical Analysis*, 33(1) :1–16, 1996.

[37] R. Herbin, W. Kheriji, and J.-C. Latché. On some implicit and semi-implicit staggered schemes for the shallow water equations. *ESAIM :M2AN*, 48 :1807–1857, 2014.

[38] S. Jin. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM Journal on Scientific Computing*, 21(2) :441–454, 1999.

[39] R. J. LeVeque. *Finite volume methods for hyperbolic problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2002.

[40] M.-S. Liou. A sequel to ausm : Ausm+. *Journal of Computational Physics*, 129(2) :364–382, 1996.

[41] M.-S. Liou and C. J. Steffen. A new flux splitting scheme. *Journal of Computational Physics*, 107(1) :23–39, 1993.

[42] X. Liu, A. Chertock, and A. Kurganov. An asymptotic preserving scheme for the two-dimensional shallow water equations with Coriolis forces. *Journal of Computational Physics*, 391 :259–279, 2019.

[43] M. Parisot and J.-P. Vila. Centered-potential regularization for the advection upstream splitting method. *SIAM Journal on Numerical Analysis*, 54(5) :3083–3104, 2015.

[44] R. Rannacher. On chorin's projection method for the incompressible navier-stokes equations. In J. G. Heywood, K. Masuda, R. Rautmann, and V. A. Solonnikov, editors, *The Navier-Stokes Equations II — Theory and Numerical Methods*, pages 167–183, Berlin, Heidelberg, 1992. Springer Berlin Heidelberg.

[45] T. Ringler, J. Thuburn, J. Klemp, and W. Skamarock. A unified approach to energy conservation and potential vorticity dynamics for arbitrarily-structured c-grids. *Journal of Computational Physics*, 229(9) :3065–3090, 2010.

[46] A. Saint-Venant. Théorie du mouvement non permanent des eaux, avec application aux crues des rivières et a l'introduction de marées dans leurs lits. *Comptes Rendus des Séances de Académie des Sciences, 73, 147, 237.*, 1871.

[47] R. Temam. Une méthode d'approximation de la solution des équations de navier-stokes. *Bulletin de la Société Mathématique de France*, 96 :115–152, 1968.

[48] J. Thuburn, T. Ringler, W. Skamarock, and J. Klemp. Numerical representation of geostrophic modes on arbitrarily structured c-grids. *Journal of Computational Physics*, 228(22) :8321–8335, 2009.

[49] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics.* Springer, 2009.

[50] R. A. Walters, E. Hanert, J. Pietrzak, and D. Le Roux. Comparison of unstructured, staggered grid methods for the shallow water equations. *Ocean Modelling*, 28(1) :106–117, 2009. The Sixth International Workshop on Unstructured Mesh Numerical Modelling of Coastal, Shelf and Ocean Flows.

[51] R. A. Walters, E. M. Lane, and E. Hanert. Useful time-stepping methods for the coriolis term in a shallow water model. *Ocean Modelling*, 28(1) :66–74, 2009. The Sixth International Workshop on Unstructured Mesh Numerical Modelling of Coastal, Shelf and Ocean Flows.

[52] H. Zakerzadeh. The rs-imex scheme for the rotating shallow water equations with the coriolis force. In C. Cances and P. Omnes, editors, *Finite Volumes for Complex Applications VIII - Methods and Theoretical Aspects*, pages 199–207. Springer, 2017.

[53] V. Zeitlin. *Geophysical fluid dynamics : understanding (almost) everything with rotating shallow water models.* Oxford University Press, 2018.

Résumé : Nous étudions dans cette thèse des méthodes numériques pour les écoulements en eaux peu profondes à surface libre. D'une part, nous nous intéressons aux modèles non-hydrostatique de Saint-Venant prenant en compte des effets dispersifs. D'autre part, nous étudions le modèle de Saint-Venant avec terme source de Coriolis et ses équilibres géostrophiques.

Dans un premier temps nous considérons des méthodes numériques pour une famille de modèles d'Euler moyennés sur la hauteur d'eau provenant de la littérature. Il s'agit de modèles de Saint-Venant avec une pression non-hydrostatique. Les méthodes numériques considérées sont basées sur une méthode de prédiction-correction consistant à décomposer le problème en deux étapes à chaque itération en temps. L'étape de prédiction mène à résoudre un système de Saint-Venant qui est habituellement résolu par une méthode de volumes finis, alors que l'étape de correction mène à résoudre un problème elliptique régissant la pression non-hydrostatique. Notre but premier est d'analyser la convergence d'un schéma mixte avec condensation de masse introduit dans la littérature en utilisant la Méthode de Discrétisation du Gradient (GDM) qui forme un cadre englobant des schémas de discrétisation usuels et récents pour des problèmes de diffusion. Cette méthode nous permet d'obtenir la convergence du schéma proposé pour le problème elliptique. Par la suite, nous proposons une nouvelle formulation conforme du problème en pression et donnons les estimateurs d'erreur correspondant via la GDM. Enfin, nous donnons un exemple d'application avec la méthode des éléments finis $\mathbb{P}_1$.

Dans un second temps nous souhaitons avoir des schémas explicites colocalisés volumes finis pour les équations de Saint-Venant non-linéaires avec un terme source de Coriolis qui seraient précis aux alentours de l'équilibre géostrophique et stables dans le cadre non-linéaire. Nous construisons plusieurs schémas volumes finis et étudions les deux propriétés suivantes : la décroissance de l'énergie semi-discrète et la préservation de l'équilibre géostrophique par la version linéarisée. Nous proposons également une version conservative de l'un de nos schéma. Enfin, nous observons le comportement de ces schémas à travers plusieurs cas tests et obtenons de meilleurs résultats en comparaison d'un schéma volumes finis classique.

Mots-clés : écoulements à surface libre, équations de Saint-Venant, pression non-hydrostatique, méthode de discrétisation du gradient, éléments finis, force de Coriolis, équilibre géostrophique, schéma équilibre, diffusion numérique, volumes finis.

Abstract : In this work we study some numerical methods for free surface shallow water flows. On one hand, we are interested in non-hydrostatic shallow water models which take into account dispersive effects. On the other hand, we study the shallow water model with Coriolis source term and the associated geostrophic equilibrium.

First, we consider numerical methods for a family of depth-averaged Euler models from the literature. These models are shallow water models with non-hydrostatic pressure. The considered numerical methods are based on a prediction-correction method which leads to decompose the problem into two parts at each time step. The prediction part consists in solving a shallow water system which is usually solved by finite volume methods, while the correction part consists in solving an elliptic problem governing the non-hydrostatic pressure. Our first goal is to analyse the convergence of a mixed mass-lumped scheme introduced in the literature through the Gradient Discretisation Method (GDM) which is a framework comprising classic and recent discretisation schemes for diffusion problems. This method allows us to obtain a convergence result on the proposed scheme for the elliptic problem. Then, we propose a new conforming formulation of the problem on the pressure and give the corresponding error estimate provided by the GDM. Finally, an application to the $\mathbb{P}_1$ conforming finite element method is given.

Then, we aim at designing explicit colocated finite volume schemes for the nonlinear Shallow Water equations with Coriolis source term that are proved to be accurate around the geostrophic equilibrium and stable in the nonlinear framework. We define some finite volume schemes and study the two properties we are interested in : the decrease of the semi-discrete energy and the preservation of the geostrophic equilibrium in the linearized version. We also propose a conservative version of one of our scheme. Finally, we illustrate the behaviour of the schemes for some standard test cases and we exhibit a great improvement when compared to a classic finite volume scheme.

Keywords : free surface flows, shallow water equations, non-hydrostatic pressure, Gradient Discretisation Method, finite element, Coriolis force, geostrophic equilibrium, well-balanced scheme, numerical diffusion, finite volume.