UPMC–M2R Parcours ANEDP
Mathématiques de la Modélisation
Méthodes de Galerkine Discontinues et Applications

Alexandre Ern[1]

22 mars 2007

L'objet de ce cours est d'approfondir l'étude de l'approximation d'EDPs hyperboliques et elliptiques stationnaires par les méthodes de Galerkine Discontinues (GD). Les problèmes modèles seront une équation d'advection–réaction dans le cas hyperbolique et une équation de diffusion–réaction dans le cas elliptique. En guise de complément, une analyse unifiée des méthodes de GD *via* la notion de systèmes de Friedrichs sera esquissée. Enfin, quelques repères bibliographiques sont fournis en fin de document (la littérature sur les méthodes de GD est particulièrement prolifique depuis quelques années).

1 Advection–réaction

Soit Ω un domaine borné et connexe de $\mathbb{R}^d$ de frontière $\partial\Omega$ lipschitzienne. Soit $f \in L^2(\Omega)$, $\mu \in L^\infty(\Omega)$ et $\beta \in [W^{1,\infty}(\Omega)]^d$ (l'hypothèse de régularité sur le gradient de β peut être considérablement affaiblie, tant que $\nabla\cdot\beta \in L^\infty(\Omega)$). On suppose que presque partout dans Ω,

$$\mu - \tfrac{1}{2}\nabla\cdot\beta \geq \mu_0 > 0. \tag{1}$$

On désigne par $\partial\Omega^\pm$ les parties entrante et sortante de la frontière de Ω ; celles-ci sont définies comme suit :

$$\partial\Omega^\pm = \{x \in \partial\Omega;\ \pm\beta(x)\cdot n(x) > 0\}, \tag{2}$$

où $n(x)$ désigne la normale extérieure à Ω au point $x \in \partial\Omega$. Le problème modèle consiste à chercher une fonction $u : \Omega \to \mathbb{R}$ telle que

$$\begin{cases} \mu u + \beta\cdot\nabla u = f & \text{dans } \Omega, \\ u = 0 & \text{sur } \partial\Omega^-. \end{cases} \tag{3}$$

La section 1.1 précise (sans rentrer dans tous les détails techniques) le cadre mathématique dans lequel formuler le problème modèle (3). Puis, la section 1.2 décrit l'approximation de ce problème par une méthode de GD et en effectue l'analyse de convergence.

1.1 Cadre mathématique

La première chose à faire est de préciser l'espace fonctionnel dans lequel on cherche la solution de (3). Pour cela, on introduit l'espace du graphe

$$W = \{v \in L^2(\Omega);\ \beta\cdot\nabla v \in L^2(\Omega)\}. \tag{4}$$

[1] CERMICS, Ecole des Ponts, ParisTech, 6 et 8, avenue Blaise Pascal, 77455 Marne la Vallée cedex 2, France – ern(at)cermics.enpc.fr

L'hypothèse $\beta\cdot\nabla v \in L^2(\Omega)$ signifie que la forme linéaire

$$\mathcal{D}(\Omega) \ni \varphi \longmapsto -\int_\Omega v\nabla\cdot(\beta\varphi) \in \mathbb{R} \tag{5}$$

est continue sur $L^2(\Omega)$; l'objet $\beta\cdot\nabla v$ désigne alors la fonction de $L^2(\Omega)$ qui peut être associée à cette forme linéaire par le biais du théorème de représentation de Riesz–Fréchet.

Lemme 1. *Equipé du produit scalaire* $(v,w)_W = (v,w)_{L^2} + (\beta\cdot\nabla v, \beta\cdot\nabla w)_{L^2}$, W *est un espace de Hilbert.*

Preuve. Il faut vérifier la complétude. Soit $(v_n)_{n\in\mathbb{N}}$ une suite de Cauchy de W. Les suites $(v_n)_{n\in\mathbb{N}}$ et $(\beta\cdot\nabla v_n)_{n\in\mathbb{N}}$ sont donc de Cauchy dans $L^2(\Omega)$. Désignons par v et w leurs limites respectives dans $L^2(\Omega)$. On a

$$-\int_\Omega v\nabla\cdot(\beta\varphi) \longleftarrow -\int_\Omega v_n\nabla\cdot(\beta\varphi) = \int_\Omega (\beta\cdot\nabla v_n)\varphi \longrightarrow \int_\Omega w\varphi,$$

ce qui montre que $v \in W$ avec $\beta\cdot\nabla v = w$ et que la suite $(v_n)_{n\in\mathbb{N}}$ tend vers v dans W. □

Le deuxième point important est de préciser le sens de la condition aux limites dans (3). On introduit l'espace

$$L^2(\partial\Omega;|\beta\cdot n|) = \{v \text{ mesurable sur } \partial\Omega;\ \textstyle\int_{\partial\Omega} |\beta\cdot n| v^2 < +\infty\}. \tag{6}$$

On suppose que

(h1) l'espace $\mathfrak{C}^1_0(\mathbb{R}^d)$ des fonctions de classe $\mathfrak{C}^1$ sur $\mathbb{R}^d$ à support compact est dense dans W ;

(h2) les frontières entrante et sortante de Ω sont bien séparées, ce qui signifie que

$$\operatorname{dist}(\partial\Omega^-, \partial\Omega^+) > 0. \tag{7}$$

On peut montrer par troncature et régularisation que l'hypothèse (h1) est satisfaite dans le cadre des hypothèses ci-dessus. L'hypothèse (h2) est plus restrictive puisqu'elle limite le champ d'application à des problèmes de transport advectif–réactif dans des tubes. Les hypothèses (h1)–(h2) permettent de montrer le théorème de trace suivant.

Lemme 2.

(i) *L'opérateur de trace*

$$\gamma : \mathfrak{C}^1_0(\mathbb{R}^d) \ni v \longmapsto v|_{\partial\Omega} \in L^2(\partial\Omega;|\beta\cdot n|) \tag{8}$$

s'étend continûment à W.

(ii) *On a la formule d'intégration par parties*

$$\forall (v,w) \in W\times W, \qquad \int_\Omega [(\beta\cdot\nabla v)w + (\beta\cdot\nabla w)v + (\nabla\cdot\beta)vw] = \int_{\partial\Omega} (\beta\cdot n)vw. \tag{9}$$

Preuve. De l'hypothèse (h2) on déduit l'existence de deux fonctions ψ^- et ψ^+ dans $\mathfrak{C}_0^1(\mathbb{R}^d)$ telles que

$$\psi^- + \psi^+ \equiv 1 \text{ dans } \Omega, \quad \psi^-|_{\partial\Omega^+} = 0, \quad \psi^+|_{\partial\Omega^-} = 0.$$

Soit $v \in \mathfrak{C}_0^1(\mathbb{R}^d)$. On a

$$\begin{aligned}\int_{\partial\Omega} v^2|\beta\cdot n| &= \int_{\partial\Omega} v^2(\psi^- + \psi^+)|\beta\cdot n| = \int_{\partial\Omega^-} v^2\psi^-|\beta\cdot n| + \int_{\partial\Omega^+} v^2\psi^+|\beta\cdot n| \\ &= -\int_{\partial\Omega} v^2\psi^-(\beta\cdot n) + \int_{\partial\Omega} v^2\psi^+(\beta\cdot n) \\ &= -\int_{\Omega} \nabla\cdot(v^2\psi^-\beta) + \int_{\Omega} \nabla\cdot(v^2\psi^+\beta) \\ &\leq c(\psi^-, \psi^+, \beta)\|v\|_W^2.\end{aligned}$$

On conclut par densité (hypothèse (h1)) pour établir le point (i). Le point (ii) devient alors immédiat : la formule (9) est vraie sur $\mathfrak{C}_0^1(\mathbb{R}^d)$ et on conclut à nouveau par densité. □

Remarque 3. *L'hypothèse de séparation des frontières entrante et sortante ne peut pas être contournée pour établir un théorème de trace dans $L^2(\partial\Omega; |\beta\cdot n|)$. Voici un contre-exemple : considérer le domaine triangulaire*

$$\Omega = \{(x, y) \in \mathbb{R}^2;\, 0 < y < 1;\, |x| < y\}$$

et le champ advectif $\beta = (1, 0)^t$. La fonction $u(x, y) = y^\alpha$ est dans W dès que $\alpha > -1$ mais elle n'admet de trace dans $L^2(\partial\Omega; |\beta\cdot n|)$ que si $\alpha > -\frac{1}{2}$.

On définit sur $W \times W$ la forme bilinéaire

$$a(v, w) = \int_{\Omega} [\mu vw + (\beta\cdot\nabla v)w] + \int_{\partial\Omega} (\beta\cdot n)^- vw, \tag{10}$$

où pour un réel $x \in \mathbb{R}$, on désigne ses parties positive et négative par

$$x^+ = \tfrac{1}{2}(|x| + x), \qquad x^- = \tfrac{1}{2}(|x| - x). \tag{11}$$

Observer que la forme bilinéaire a est bien continue sur $W \times W$ de par le lemme 2. De plus, cette forme bilinéaire est L^2-coercive sur W. C'est une conséquence immédiate de l'hypothèse (1) et de la formule d'intégration par parties (9) puisque

$$\begin{aligned}a(v, v) &= \int_{\Omega} (\mu - \tfrac{1}{2}\nabla\cdot\beta)v^2 + \tfrac{1}{2}\int_{\partial\Omega} (\beta\cdot n)v^2 + \int_{\partial\Omega} (\beta\cdot n)^- v^2 \\ &\geq \mu_0\|v\|_{L^2}^2 + \tfrac{1}{2}\int_{\partial\Omega} |\beta\cdot n|v^2.\end{aligned} \tag{12}$$

On considère le problème suivant :

$$\text{Chercher } u \in W \text{ tel que } a(u, w) = \int_{\Omega} fw,\ \forall w \in W. \tag{13}$$

La L^2-coercivité de la forme bilinéaire a nous assure de l'unicité de la solution de ce problème. Nous n'établirons pas ici l'existence, mais admettrons le résultat suivant.

Théorème 4. *Le problème* (13) *est bien posé. De plus, son unique solution est solution de* (3).

1.2 Approximation GD

L'ojet de cette section est l'analyse numérique d'une approximation de type GD de la solution du problème (13). En particulier, on s'attachera à décrire précisément les étapes (tout à fait naturelles) dans la formulation du problème discret. Mais auparavant, introduisons quelques notations utiles.

1.2.1 Notations

Soit $\{\mathcal{T}_h\}_{h>0}$ une famille régulière (au sens usuel) de maillages affines du domaine Ω. On suppose pour simplifier que Ω est un polygone ou un polyèdre de $\mathbb{R}^d$ afin que chaque maillage $\mathcal{T}_h$ recouvre exactement Ω. Sans perte de généralité, on suppose que $h \leq 1$. Pour tout entier $s \geq 1$, $H^s(\mathcal{T}_h)$ désigne l'espace de Sobolev brisé (dont les élements sont des fonctions dont la restriction à chaque maille $T \in \mathcal{T}_h$ est dans $H^s(T)$). Il sera commode de disposer de l'opérateur de gradient discret $\nabla_h : H^1(\mathcal{T}_h) \to [L^2(\Omega)]^d$ tel que pour $v \in H^1(\mathcal{T}_h)$, $(\nabla_h v)|_T = \nabla(v|_T)$ pour tout $T \in \mathcal{T}_h$, ainsi que de l'opérateur de divergence discrète $\nabla_h \cdot : [H^1(\mathcal{T}_h)]^d \to L^2(\Omega)$ tel que pour $\sigma \in [H^1(\mathcal{T}_h)]^d$, $(\nabla_h \cdot \sigma)|_T = \nabla \cdot (\sigma|_T)$ pour tout $T \in \mathcal{T}_h$.

On désigne par $\mathcal{F}_h$ l'ensemble des faces du maillage. Cet ensemble est partitionné en $\mathcal{F}_h = \mathcal{F}_h^{\mathrm{i}} \cup \mathcal{F}_h^{\partial}$ où $\mathcal{F}_h^{\mathrm{i}}$ désigne l'ensemble des faces intérieures du maillage et $\mathcal{F}_h^{\partial}$ l'ensemble des faces situées sur la frontière $\partial\Omega$. Pour $F \in \mathcal{F}_h^{\mathrm{i}}$, il existe deux mailles T_1 et T_2 dans $\mathcal{T}_h$ telles que $F = T_1 \cap T_2$. On pose $\Delta_F = T_1 \cup T_2$ et on désigne par n_F le vecteur unitaire normal à F pointant de T_1 vers T_2. Pour $v \in H^1(\mathcal{T}_h)$, on définit son saut à travers F et sa moyenne sur F par les formules

$$[\![v]\!]_F = v_1 - v_2, \qquad \{v\}_F = \tfrac{1}{2}(v_1 + v_2), \tag{14}$$

où pour alléger les notations, on a posé $v_i = v|_{T_i}$, $i \in \{1, 2\}$. L'indice F est omis dans les sauts et les moyennes lorsqu'il n'y a pas d'ambiguïté. Pour $F \in \mathcal{F}_h^{\partial}$, on pose $\Delta_F = T$ où T est l'unique maille de $\mathcal{T}_h$ dont F est une face, n_F désigne la normale unitaire sur F sortante de Ω et les saut et valeur moyenne de $v \in H^1(\mathcal{T}_h)$ sont pris par convention égaux à la restriction de v sur F.

Par la suite, la notation $A \lesssim B$ signifie qu'il existe une constante positive c, *indépendante du maillage*, telle que $A \leq cB$. Pour alléger les notations, on désigne par $(\cdot, \cdot)_E$ le produit scalaire usuel de $L^2(E)$ (où E est une maille, une face du maillage ou une union de tels objets) et par $\|\cdot\|_E$ la norme associée.

Soit un entier $p \geq 1$. On introduit l'espace de Galerkine Discontinu

$$\mathbb{G}_h^p = \{v_h \in L^2(\Omega); \forall T \in \mathcal{T}_h,\, v_h|_T \in \mathbb{P}_p\}, \tag{15}$$

où $\mathbb{P}_p$ désigne l'espace vectoriel des polynômes de degré total inférieur ou égal à p. La dimension de $\mathbb{G}_h^p$ est égale à $N_{\mathrm{ma}} C_{p+d}^p$ où N_{ma} désigne le nombre de mailles de $\mathcal{T}_h$ et C_{p+d}^p le coefficient du binôme de Pascal associé aux indices $(p+d)$ et p (on rappelle que C_{p+d}^p est égal à la dimension de l'espace vectoriel $\mathbb{P}_p$). L'espace $\mathbb{G}_h^p$ possède des propriétés d'interpolation optimales dans les espaces de Sobolev brisés. Ainsi, pour tout $z \in H^{p+1}(\mathcal{T}_h)$, il existe $z_h \in \mathbb{G}_h^p$ tel que pour tout $T \in \mathcal{T}_h$,

$$\|z - z_h\|_T + h_T^{\frac{1}{2}} \|z - z_h\|_{\partial T} + h_T \|\nabla(z - z_h)\|_T \lesssim h_T^{p+1} \|z\|_{H^{p+1}(T)}. \tag{16}$$

Enfin, grâce à la régularité de la famille $\{\mathcal{T}_h\}_{h>0}$, on dispose dans $\mathbb{G}_h^p$ des inégalités inverse et de trace suivantes : $\forall v_h \in \mathbb{G}_h^p$, $\forall T \in \mathcal{T}_h$,

$$\|\nabla v_h\|_T \lesssim h_T^{-1}\|v_h\|_T, \qquad \|v_h\|_{\partial T} \lesssim h_T^{-\frac{1}{2}}\|v_h\|_T. \tag{17}$$

1.2.2 Formulation du problème discret

On pose $W_h := \mathbb{G}_h^p$. Le problème discret consiste à

$$\text{Chercher } u_h \in W_h \text{ tel que } a_h(u_h, w_h) = \int_\Omega f w_h,\ \forall w_h \in W_h, \tag{18}$$

et il s'agit maintenant de construire la forme bilinéaire discrète a_h. Deux hypothèses de régularité supplémentaires sont faites sur la solution exacte :

- $u \in H^{p+1}(\mathcal{T}_h)$, ce qui signifie que le maillage a été construit de sorte que la solution exacte soit suffisamment régulière à l'intérieur de chaque maille ;
- pour tout $F \in \mathcal{F}_h^{\mathrm{i}}$, $\beta{\cdot}n_F[\![u]\!] = 0$, ce qui implique que la solution exacte ne peut sauter à travers une face que si le champ advectif est tangent à cette face.

Pour que ces hypothèses soient satisfaites lorsque la solution exacte présente des singularités, il faut en général adapter le maillage à ces singularités. La figure 1 propose un exemple d'une telle situation. Le champ d'advection est représenté à gauche ; il est supposé uniforme. On suppose de plus que sur le bord latéral gauche du domaine (ici, un carré), on impose une condition aux limites de Dirichlet non homogène avec une discontinuité au point S. Dans ces conditions, la solution exacte est discontinue sur la ligne de courant traversant le domaine et représentée en gras sur la figure. On constate que lorsque le maillage est adapté à cette singularité (cas à gauche), les deux hypothèses de régularité ci-dessus sont raisonnables ; dans la situation représentée à droite, ces hypothèses tombent clairement en défaut.

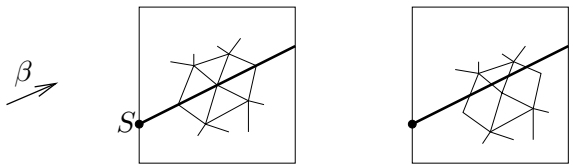


FIG. 1 – Maillage adapté à la singularité de la solution exacte à gauche ; maillage inadapté à droite.

Les deux propriétés essentielles de la forme bilinéaire a_h sur lesquelles repose toute la construction ci-dessous sont les suivantes :

(i) L^2-coercivité sur W_h : il s'agit du pendant discret de la propriété (12), à savoir

$$\forall v_h \in W_h, \qquad a_h(v_h, v_h) \gtrsim \|v_h\|_{L^2}^2 + \int_{\partial\Omega} |\beta{\cdot}n| v_h^2. \tag{19}$$

(ii) Consistance forte, à savoir

$$a_h(u, w_h) = \int_\Omega f w_h, \quad \forall w_h \in W_h, \tag{20}$$

où $u \in W$ désigne la solution exacte, c'est-à-dire la solution du problème (13).

Le choix le plus simple consiste à prendre $a_h := a$ en localisant le gradient sur chaque élément,

$$a_h(v_h, w_h) = \int_\Omega [\mu v_h w_h + (\beta{\cdot}\nabla_h v_h) w_h] + \int_{\partial\Omega} (\beta{\cdot}n)^- v_h w_h, \tag{21}$$

car cela assure *de facto* la consistance. Toutefois, ce choix n'est pas compatible avec notre exigence de L^2-coercivité discrète comme le montre le calcul suivant : pour tout $v_h \in V_h$,

$$\begin{aligned} a(v_h, v_h) &= \int_\Omega [\mu v_h^2 + (\beta{\cdot}\nabla_h v_h) v_h] + \int_{\partial\Omega} (\beta{\cdot}n)^- v_h^2 \\ &= \int_\Omega (\mu - \tfrac{1}{2}\nabla{\cdot}\beta) v_h^2 + \tfrac{1}{2}\int_{\partial\Omega} |\beta{\cdot}n| v_h^2 + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F \{\beta{\cdot}n v_h^2\}. \end{aligned}$$

Or, avec des notations évidentes,

$$\{\beta{\cdot}n v_h^2\} = \tfrac{1}{2}(\beta_1{\cdot}n_1 v_1^2 + \beta_2{\cdot}n_2 v_2^2) = \beta{\cdot}n_F [\![v_h]\!]\{v_h\}.$$

(On rappelle que par hypothèse, le champ advectif β est continu à travers F.) Par conséquent, comme la solution exacte vérifie par hypothèse $\beta{\cdot}n_F[\![u]\!] = 0$ pour tout $F \in \mathcal{F}_h^{\mathrm{i}}$, une modification consistante du choix (21) qui rétablit la L^2-coercivité discrète consiste à poser

$$a_h(v_h, w_h) = \int_\Omega [\mu v_h w_h + (\beta{\cdot}\nabla_h v_h) w_h] + \int_{\partial\Omega} (\beta{\cdot}n)^- v_h w_h - \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F \beta{\cdot}n_F [\![v_h]\!]\{w_h\}. \tag{22}$$

Il sera commode de raisonner sur la formule équivalente obtenue en intégrant par parties la dérivée advective,

$$a_h(v_h, w_h) = \int_\Omega [(\mu - \nabla{\cdot}\beta) v_h w_h - v_h(\beta{\cdot}\nabla_h w_h)] + \int_{\partial\Omega} (\beta{\cdot}n)^+ v_h w_h + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F \beta{\cdot}n_F \{v_h\}[\![w_h]\!]. \tag{23}$$

Cette formule résulte de l'identité

$$\begin{aligned} \int_\Omega (\beta{\cdot}\nabla_h v_h) w_h = &- \int_\Omega [(\nabla{\cdot}\beta) v_h w_h - v_h(\beta{\cdot}\nabla_h w_h)] + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F 2\{\beta{\cdot}n v_h w_h\} \\ &+ \sum_{F\in\mathcal{F}_h^{\partial}} \int_F \beta{\cdot}n v_h w_h, \end{aligned}$$

et du fait que

$$2\{\beta{\cdot}n v_h w_h\} = \beta{\cdot}n_F(\{v_h\}[\![w_h]\!] + [\![v_h]\!]\{w_h\}).$$

La forme bilinéaire a_h définie par (22) ou (23) est d'une part consistante et d'autre part elle satisfait la propriété de L^2-coercivité discrète (19).

Avant de poursuivre la construction de la forme bilinéaire a_h (qui n'est pas encore terminée !), arrêtons-nous un instant pour considérer le problème discret (18) obtenu avec le choix (23) pour la forme bilinéaire a_h. Ce problème admet une reformulation locale obtenue en considérant une fonction test de la forme $w_h = q1_T$ où q est un polynôme

arbitraire dans $\mathbb{P}_p$ et où 1_T désigne la fonction indicatrice de la maille T. Ainsi, u_h est solution de (18) si et seulement si pour tout maille $T \in \mathcal{T}_h$ et pour tout polynôme $q \in \mathbb{P}_p$,

$$\int_T [(\mu - \nabla\cdot\beta)u_h q - u_h(\beta\cdot\nabla q)] + \sum_{F\subset\partial T} \int_F \phi_{T,F}(u_h) q = \int_T fq, \tag{24}$$

où le *flux* $\phi_{T,F}(u_h)$ est défini par

$$\phi_{T,F}(u_h) = \begin{cases} \beta\cdot n_T \{u_h\}, & F \in \mathcal{F}_h^{\mathrm{i}}, \\ (\beta\cdot n)^+ u_h, & F \in \mathcal{F}_h^{\partial}, \end{cases} \tag{25}$$

n_T désignant la normale unitaire extérieure à la maille T. Lorsqu'on prend pour q une fonction constante égale à 1, on obtient une équation de bilan de masse locale sous la forme

$$\int_T (\mu - \nabla\cdot\beta)u_h + \sum_{F\subset\partial T} \int_F \phi_{T,F}(u_h) = \int_T f. \tag{26}$$

Le flux $\phi_{T,F}(u_h)$ est appelé «flux centré» du fait de l'utilisation de la valeur moyenne de u_h pour $F \in \mathcal{F}_h^{\mathrm{i}}$. Notons deux propriétés importantes de ce flux.

1. Conservativité : si $F = T_1 \cap T_2$,

$$\phi_{T_1,F}(u_h) + \phi_{T_2,F}(u_h) = 0. \tag{27}$$

   Cela signifie que ce qui «sort» par une face $F \in \mathcal{F}_h^{\mathrm{i}}$ d'une maille «rentre» dans la maille voisine.

2. Consistance avec les flux exacts. Ceux-ci sont obtenus formellement en intégrant par parties sur une maille $T \in \mathcal{T}_h$ l'EDP $\mu u + \beta\cdot\nabla u = f$ multipliée par un polynôme $q \in \mathbb{P}_p$, ce qui donne

$$\int_T [(\mu - \nabla\cdot\beta)uq - u(\beta\cdot\nabla q)] + \sum_{F\subset\partial T} \int_F \Phi_{T,F}(u) q = \int_T fq, \tag{28}$$

   avec le flux exact

$$\Phi_{T,F}(u) = \begin{cases} (\beta\cdot n_T)u, & F \in \mathcal{F}_h^{\mathrm{i}}, \\ (\beta\cdot n)u, & F \in \mathcal{F}_h^{\partial}. \end{cases} \tag{29}$$

   En utilisant la condition aux limites $u|_{\partial\Omega^-} = 0$ et la propriété $\beta\cdot n_F[\![u]\!] = 0$ pour tout $F \in \mathcal{F}_h^{\mathrm{i}}$, il vient

$$\phi_{T,F}(u) = \Phi_{T,F}(u). \tag{30}$$

Reprenons maintenant la construction de la forme bilinéaire a_h. Les fonctions de W_h sont *a priori* discontinues de maille à maille ; or, nous avons vu que la solution exacte satisfait $\beta\cdot n_F[\![u]\!] = 0$ pour tout $F \in \mathcal{F}_h^{\mathrm{i}}$. Ceci plaide pour une modification de la forme bilinéaire a_h en y ajoutant une pénalisation au sens des moindres carrés des sauts aux interfaces. On considère

$$\begin{aligned} a_h(v_h, w_h) = &\int_\Omega [\mu v_h w_h + (\beta\cdot\nabla_h v_h) w_h] + \int_{\partial\Omega} (\beta\cdot n)^- v_h w_h - \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F \beta\cdot n_F [\![v_h]\!]\{w_h\} \\ &+ \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F \frac{\alpha}{2} |\beta\cdot n_F| [\![v_h]\!][\![w_h]\!], \end{aligned} \tag{31}$$

ou encore, en intégrant par parties la dérivée advective,

$$\begin{aligned} a_h(v_h,w_h) = &\int_\Omega [(\mu - \nabla\cdot\beta)v_h w_h - v_h(\beta\cdot\nabla_h w_h)] + \int_{\partial\Omega} (\beta\cdot n)^+ v_h w_h \\ &+ \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F \beta\cdot n_F \{v_h\}[\![w_h]\!] + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F \frac{\alpha}{2}|\beta\cdot n_F|[\![v_h]\!][\![w_h]\!], \end{aligned} \tag{32}$$

où $\alpha > 0$ est un paramètre choisi par le numéricien (celui-ci peut éventuellement être pris différent sur chaque face). Il est important d'observer que cette nouvelle forme bilinéaire a_h satisfait bien les propriétés de L^2-coercivité sur W_h et de consistance. On a en effet

$$a_h(v_h,v_h) \geq \mu_0 \|v_h\|_{L^2}^2 + \sum_{F\in\mathcal{F}_h^{\partial}} \int_F \frac{1}{2}|\beta\cdot n| v_h^2 + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F \frac{\alpha}{2}|\beta\cdot n_F|[\![v_h]\!]^2, \qquad \forall v_h \in W_h, \tag{33}$$

$$a_h(u,w_h) = \int_\Omega f w_h, \qquad \forall w_h \in W_h. \tag{34}$$

On notera que la propriété de coercivité (33) est plus forte que (19) puisqu'elle comporte en plus un contrôle sur les sauts de v_h à travers les interfaces.

L'analyse de convergence présentée à la section suivante montrera que la forme bilinéaire a_h ainsi construite conduit à des estimations d'erreur (quasi-)optimales. Toutefois, avant d'aborder ce point, voyons comment la pénalisation au sens des moindres carrés des sauts aux interfaces modifie les flux. On vérifie facilement que les flux associés à la nouvelle forme bilinéaire a_h s'écrivent

$$\phi_{T,F}(u_h) = \begin{cases} \beta\cdot n_T \{u_h\} + \frac{\alpha}{2}|\beta\cdot n_T|[\![u_h]\!]_T, & F\in\mathcal{F}_h^{\mathrm{i}}, \\ (\beta\cdot n)^+ u_h, & F\in\mathcal{F}_h^{\partial}, \end{cases} \tag{35}$$

où $[\![u_h]\!]_T$ désigne le saut associé à la maille T défini comme la différence entre la valeur intérieure à T moins la valeur extérieure à T ; on a $[\![u_h]\!]_T = n_T\cdot n_F [\![u_h]\!]_F$. Les flux ci-dessus sont conservatifs et consistants avec ceux de la solution exacte. Un cas particulier remarquable est celui pour lequel $\alpha = 1$. On obtient

$$\phi_{T,F}(u_h) = \begin{cases} \beta\cdot n_T u_h^{\uparrow}, & F\in\mathcal{F}_h^{\mathrm{i}}, \\ (\beta\cdot n)^+ u_h, & F\in\mathcal{F}_h^{\partial}, \end{cases} \tag{36}$$

où $u_h^{\uparrow}$ désigne la valeur intérieure à T si $\beta\cdot n_T > 0$ et la valeur extérieure sinon. Ce flux est connu sous le nom de «flux upwind» car il consiste à choisir sur une interface la valeur de u_h située en amont des lignes de courant associées à β. Une formule équivalente consiste à écrire

$$\phi_{T,F}(u_h) = \begin{cases} (\beta\cdot n_T)^+ u_h^{\mathrm{int}} + (\beta\cdot n_T)^- u_h^{\mathrm{ext}}, & F\in\mathcal{F}_h^{\mathrm{i}}, \\ (\beta\cdot n)^+ u_h, & F\in\mathcal{F}_h^{\partial}, \end{cases} \tag{37}$$

avec des notations évidentes.

1.2.3 Analyse de convergence

Afin de bien comprendre comment s'articule l'analyse de convergence, reprenons-en les étapes essentielles dans un cadre général. La solution discrète u_h vit dans l'espace W_h et la solution exacte u dans l'espace W. Comme l'approximation n'est pas conforme ($W_h \not\subset W$), on est amené à considérer l'espace vectoriel somme $Z := W + W_h$ dans lequel vit l'erreur $u - u_h$. La première propriété à établir est une propriété de stabilité ; celle-ci prend en général la forme d'une condition inf-sup discrète faisant intervenir une norme $|||\cdot|||$:

$$\forall v_h \in W_h, \qquad |||v_h||| \lesssim \sup_{w_h \in W_h \setminus \{0\}} \frac{a_h(v_h, w_h)}{|||w_h|||}. \tag{38}$$

C'est par le biais de cette norme que sera estimée l'erreur $u - u_h$; cette norme doit donc être définie sur Z. Evidemment, si on dispose d'une propriété de coercivité sur W_h pour la forme bilinéaire a_h (c'est-à-dire, si on sait que $|||v_h|||^2 \lesssim a_h(v_h, v_h)$ pour tout $v_h \in W_h$), cela suffit à établir une condition inf-sup discrète, puisque

$$|||v_h||| \lesssim \frac{a_h(v_h, v_h)}{|||v_h|||} \leq \sup_{w_h \in W_h \setminus \{0\}} \frac{a_h(v_h, w_h)}{|||w_h|||}.$$

Toutefois, la norme $|||\cdot|||$ dans la condition inf-sup doit être suffisamment grande pour assurer également une propriété de continuité sous la forme

$$\forall z \in Z,\ \forall w_h \in W_h, \qquad a_h(z, w_h) \lesssim |||z|||_* |||w_h|||, \tag{39}$$

où $|||\cdot|||_*$ est une norme définie sur Z et qui doit contrôler la norme $|||\cdot|||$. Dans ces conditions, on a en effet pour tout $y_h \in W_h$,

$$|||u_h - y_h||| \lesssim \sup_{w_h \in W_h \setminus \{0\}} \frac{a_h(u_h - y_h, w_h)}{|||w_h|||} = \sup_{w_h \in W_h \setminus \{0\}} \frac{a_h(u - y_h, w_h)}{|||w_h|||},$$

grâce à la propriété de consistance. Par suite, de par la propriété de continuité,

$$|||u_h - y_h||| \lesssim |||u - y_h|||_*,$$

et par inégalité triangulaire, il vient

$$\begin{aligned} |||u - u_h||| &\leq |||u - y_h||| + |||u_h - y_h||| \\ &\lesssim |||u - y_h|||_* + |||u - y_h|||_* \\ &\lesssim |||u - y_h|||_*. \end{aligned}$$

Prenant l'infimum sur $y_h \in W_h$, on aboutit à l'estimation d'erreur optimale

$$|||u - u_h||| \lesssim \inf_{y_h \in W_h} |||u - y_h|||_*. \tag{40}$$

Pour le problème d'advection-réaction, la norme de stabilité qui conduit à des estimations d'erreur optimales dans la norme du graphe brisée et des estimations d'erreur quasi-optimales en norme L^2 (suboptimalité d'ordre $\frac{1}{2}$) est la suivante :

$$|||v|||^2 = \|v\|_{L^2}^2 + |v|_M^2 + |v|_J^2 + \sum_{T \in \mathcal{T}_h} h_T \|\beta \cdot \nabla v\|_T^2, \tag{41}$$

avec

$$|v|_M^2 = \sum_{F\in\mathcal{F}_h^\partial} |v|_{M,F}^2, \qquad |v|_{M,F}^2 = \int_F |\beta\cdot n| v^2, \tag{42}$$

$$|v|_J^2 = \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} |v|_{J,F}^2, \qquad |v|_{J,F}^2 = \int_F \tfrac{\alpha}{2}|\beta\cdot n_F| [\![v]\!]^2. \tag{43}$$

Puisque la solution exacte est dans $W \cap H^{p+1}(\mathcal{T}_h)$ par hypothèse, on travaillera dans $Z := (W \cap H^{p+1}(\mathcal{T}_h)) + W_h$. Il est clair que la norme $|\!|\!| \cdot |\!|\!|$ est bien définie sur cet espace.

Lemme 5 (Stabilité)**.** *On a*

$$\forall v_h \in W_h, \qquad |\!|\!| v_h |\!|\!| \lesssim \sup_{w_h\in W_h\setminus\{0\}} \frac{a_h(v_h,w_h)}{|\!|\!| w_h |\!|\!|}. \tag{44}$$

Preuve. Soit $v_h \in W_h$. On pose $\mathbb{S} = \sup_{w_h\in W_h\setminus\{0\}} \frac{a_h(v_h,w_h)}{|\!|\!| w_h |\!|\!|}$.
(i) La propriété (33) implique que

$$\|v_h\|_{L^2}^2 + |v_h|_M^2 + |v_h|_J^2 \lesssim a_h(v_h,v_h) \lesssim \mathbb{S}|\!|\!| v_h |\!|\!|. \tag{45}$$

(ii) Il reste à contrôler la dérivée advective. On définit $\overline{\beta}$ comme le champ vectoriel constant par morceaux et égal sur chaque maille à la valeur moyenne de β sur cette maille. L'hypothèse de régularité sur β implique que

$$\forall T \in \mathcal{T}_h, \qquad \|\overline{\beta} - \beta\|_{L^\infty(T)} \lesssim h_T.$$

On considère la fonction π_h telle que pour tout $T \in \mathcal{T}_h$, $\pi_h|_T = h_T\overline{\beta}\cdot\nabla v_h$. Par construction, $\pi_h \in W_h$. De plus, en utilisant des inégalités inverses et de trace,

$$\begin{aligned}
|\!|\!| \pi_h |\!|\!|^2 &= \|\pi_h\|_{L^2}^2 + |\pi_h|_M^2 + |\pi_h|_J^2 + \sum_{T\in\mathcal{T}_h} h_T\|\beta\cdot\nabla\pi_h\|_T^2 \\
&\lesssim \|v_h\|_{L^2}^2 + \sum_{F\in\mathcal{F}_h} h_F^{-1}h_F^2\|\overline{\beta}\cdot\nabla_h v_h\|_{\Delta_F}^2 + \sum_{T\in\mathcal{T}_h} h_T h_T^{-2}h_T^2\|\overline{\beta}\cdot\nabla v_h\|_T^2 \\
&\lesssim \|v_h\|_{L^2}^2 + \sum_{T\in\mathcal{T}_h} h_T\|\beta\cdot\nabla v_h\|_T^2 \lesssim |\!|\!| v_h |\!|\!|^2,
\end{aligned}$$

puisque

$$\|\overline{\beta}\cdot\nabla v_h\|_T^2 \lesssim \|\beta\cdot\nabla v_h\|_T^2 + \|(\overline{\beta}-\beta)\cdot\nabla v_h\|_T^2 \lesssim \|\beta\cdot\nabla v_h\|_T^2 + \|v_h\|_T^2.$$

Par ailleurs,

$$\sum_{F\in\mathcal{F}_h} \|\pi_h\|_F^2 \lesssim \sum_{T\in\mathcal{T}_h} h_T^{-1}h_T^2\|\overline{\beta}\cdot\nabla v_h\|_T^2 \lesssim |\!|\!| v_h |\!|\!|^2.$$

De plus,

$$\begin{aligned}\sum_{T\in\mathcal{T}_h} h_T\|\beta\cdot\nabla v_h\|_T^2 &= a_h(v_h,\pi_h) - (\mu v_h,\pi_h)_\Omega - \sum_{T\in\mathcal{T}_h} h_T(\beta\cdot\nabla v_h,(\overline{\beta}-\beta)\cdot\nabla v_h)_T\\ &\quad - \sum_{F\in\mathcal{F}_h^\partial}\int_F(\beta\cdot n)^- v_h\pi_h + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}}\int_F \beta\cdot n_F[\![v_h]\!]\{\pi_h\}\\ &\quad - \sum_{F\in\mathcal{F}_h^{\mathrm{i}}}\int_F\frac{\alpha}{2}|\beta\cdot n_F|[\![v_h]\!][\![\pi_h]\!]\\ &:= T_1+\ldots+T_6.\end{aligned}$$

On constate que

$$|T_1|\le \mathbb{S}|\!|\!|\pi_h|\!|\!| \lesssim \mathbb{S}|\!|\!|v_h|\!|\!|.$$

De plus, en utilisant des inégalités inverses et de trace,

$$\begin{aligned}|T_2|+|T_4|+|T_5|+|T_6| &\lesssim \|v_h\|_\Omega\|\pi_h\|_\Omega + |v_h|_M\left(\sum_{F\in\mathcal{F}_h^\partial}\int_F\pi_h^2\right)^{\frac12}\\ &\quad + |v_h|_J\left(\sum_{F\in\mathcal{F}_h^{\mathrm{i}}}\int_F\{\pi_h\}^2+[\![\pi_h]\!]^2\right)^{\frac12}\\ &\lesssim \left(\|v_h\|_\Omega^2+|v_h|_M^2+|v_h|_J^2\right)^{\frac12}\left(\|\pi_h\|_\Omega^2+\sum_{F\in\mathcal{F}_h}\|\pi_h\|_F^2\right)^{\frac12}\\ &\lesssim \mathbb{S}^{\frac12}|\!|\!|v_h|\!|\!|^{\frac32}.\end{aligned}$$

Enfin,

$$|T_3|\lesssim\left(\sum_{T\in\mathcal{T}_h}h_T\|\beta\cdot\nabla v_h\|_T^2\right)^{\frac12}\left(\sum_{T\in\mathcal{T}_h}h_T h_T^2 h_T^{-2}\|v_h\|_T^2\right)^{\frac12}\lesssim\gamma\sum_{T\in\mathcal{T}_h}h_T\|\beta\cdot\nabla v_h\|_T^2+\|v_h\|_\Omega^2,$$

où γ peut être choisi aussi petit que nécessaire. En collectant les inégalités ci-dessus, on aboutit à l'estimation

$$\sum_{T\in\mathcal{T}_h}h_T\|\beta\cdot\nabla v_h\|_T^2\lesssim\mathbb{S}^{\frac12}|\!|\!|v_h|\!|\!|^{\frac32}+\mathbb{S}|\!|\!|v_h|\!|\!|. \tag{46}$$

(iii) En combinant (45) et (46), il vient

$$|\!|\!|v_h|\!|\!|^2\lesssim\mathbb{S}^{\frac12}|\!|\!|v_h|\!|\!|^{\frac32}+\mathbb{S}|\!|\!|v_h|\!|\!|,$$

d'où on déduit facilement que $|\!|\!|v_h|\!|\!|\lesssim\mathbb{S}$ en utilisant deux fois l'inégalité de Young. □

Nous allons maintenant établir une propriété de continuité pour la forme bilinéaire a_h ; la preuve en sera nettement plus simple. On introduit la norme

$$|\!|\!|v|\!|\!|_*^2=|\!|\!|v|\!|\!|^2+\sum_{T\in\mathcal{T}_h}[h_T^{-1}\|v\|_T^2+\|v\|_{\partial T}^2]. \tag{47}$$

Lemme 6 (Continuité). *On a*

$$\forall z \in Z,\ \forall w_h \in W_h, \qquad a_h(z, w_h) \lesssim |\!|\!| z |\!|\!|_* |\!|\!| w_h |\!|\!|. \tag{48}$$

Preuve. Considérons la formulation (32) et majorons les différents termes composant le membre de droite. Il vient

$$\begin{aligned}
\int_\Omega (\mu - \nabla\cdot\beta) z w_h &\lesssim \|z\|_\Omega \|w_h\|_\Omega, \\
\int_\Omega z(\beta\cdot\nabla_h w_h) &\lesssim \left(\sum_{T\in\mathcal{T}_h} h_T^{-1} \|z\|_T^2 \right)^{\frac{1}{2}} \left(\sum_{T\in\mathcal{T}_h} h_T \|\beta\cdot\nabla w_h\|_T^2 \right)^{\frac{1}{2}}, \\
\sum_{F\in\mathcal{F}_h^\partial} \int_F (\beta\cdot n)^+ z w_h &\lesssim \left(\sum_{F\in\mathcal{F}_h^\partial} \|z\|_F^2 \right)^{\frac{1}{2}} |w_h|_M, \\
\sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F \beta\cdot n_F \{z\} [\![w_h]\!] &\lesssim \left(\sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \|z\|_F^2 \right)^{\frac{1}{2}} |w_h|_J, \\
\sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F \frac{\alpha}{2} |\beta\cdot n_F| [\![z]\!] [\![w_h]\!] &\lesssim |z|_J |w_h|_J,
\end{aligned}$$

et on conclut aisément. □

Nous sommes maintenant en mesure d'énoncer le résultat principal de cette section.

Théorème 7 (Convergence). *On a*

$$|\!|\!| u - u_h |\!|\!| \lesssim h^{p+\frac{1}{2}} \|u\|_{H^{p+1}(\mathcal{T}_h)}. \tag{49}$$

Preuve. Utiliser le cadre théorique ci-dessus pour obtenir

$$|\!|\!| u - u_h |\!|\!| \lesssim \inf_{y_h\in W_h} |\!|\!| u - y_h |\!|\!|_*,$$

puis utiliser le résultat d'interpolation (16) pour estimer le membre de droite. □

En particulier, on a

$$\|u - u_h\|_\Omega \lesssim h^{p+\frac{1}{2}} \|u\|_{H^{p+1}(\mathcal{T}_h)}, \tag{50}$$

ce qui est légèrement suboptimal (d'ordre $\frac{1}{2}$), et si le maillage est quasi-uniforme,

$$\|\beta\cdot\nabla_h(u - u_h)\|_\Omega \lesssim h^p \|u\|_{H^{p+1}(\mathcal{T}_h)}, \tag{51}$$

ce qui est optimal.

2 Diffusion–réaction : approche à deux champs

Soit $f \in L^2(\Omega)$. On considère le problème modèle

$$\begin{cases} -\Delta u + u = f & \text{dans } \Omega, \\ u = 0 & \text{sur } \partial\Omega. \end{cases} \tag{52}$$

Nous avons choisi de travailler sur ce problème relativement simple afin de mettre en avant le traitement du terme de diffusion par les méthodes de GD. Les développements ci-dessous peuvent être étendus dans de nombreuses directions :

- on peut considérer d'autres conditions aux limites que celles de Dirichlet homogène, par exemple des conditions aux limites de Neumann ou de Robin ;
- on peut également considérer une équation de diffusion–advection–réaction générale sous la forme
$$-\nabla\cdot(K\nabla u) + \beta\cdot\nabla u + \mu u = f, \tag{53}$$
où le terme de réaction μ et le champ advectif β satisfont les hypothèses énoncées à la section précédente et où le tenseur de diffusion $K \in [L^\infty(\Omega)]^{d,d}$ est supposé symétrique et uniformément défini positif ;
- on peut omettre, *modulo* l'utilisation d'inégalités de type Poincaré, le terme d'ordre zéro dans (52) ;
- on peut enfin étudier dans (53) divers régimes asymptotiques dans les coefficients phénoménologiques K, β et μ, comme par exemple celui où le terme advectif domine le terme diffusif ou celui où le tenseur K présente des fortes hétérogénités ou des fortes anisotropies.

Dans cette section, le problème (52) est reformulé sous forme mixte, ce qui consiste à introduire la variable auxiliaire $\sigma = -\nabla u$ (observer que u est à valeurs dans $\mathbb{R}$ et que σ est à valeurs dans $\mathbb{R}^d$). On dit que u est la variable primale et σ le flux (diffusif). On considère donc le problème suivant :

$$\begin{cases} \sigma + \nabla u = 0 & \text{dans } \Omega, \\ \nabla\cdot\sigma + u = f & \text{dans } \Omega, \\ u = 0 & \text{sur } \partial\Omega. \end{cases} \tag{54}$$

La démarche suivie afin de construire une approximation GD de (54) est la même que pour l'équation d'advection–réaction :

1. au niveau continu, formuler le problème (54) en utilisant l'espace du graphe et une forme bilinéaire continue sur cet espace qui impose les conditions aux limites au sens faible ;
2. au niveau discret, modifier la forme bilinéaire en y rajoutant tout d'abord un terme de consistance qui préserve une propriété de L^2-coercivité discrète ; puis, pénaliser au sens des moindres carrés les sauts aux interfaces de (certaines composantes de) σ et de u ;
3. enfin, effectuer l'analyse de convergence en établissant comme précédemment une propriété de stabilité (condition inf-sup discrète) et une propriété de continuité.

2.1 Cadre mathématique

L'espace du graphe naturellement associé à (54) est

$$\begin{aligned} W &= H(\mathrm{div};\Omega)\times H^1(\Omega)\\ &= \{(\sigma,u)\in[L^2(\Omega)]^{d+1};\ \nabla\cdot\sigma\in L^2(\Omega);\ \nabla u\in[L^2(\Omega)]^d\}. \end{aligned} \tag{55}$$

L'espace W ayant une structure d'espace produit, il sera commode par la suite de décomposer un élément $z\in W$ sous la forme (z^σ,z^u) avec $z^\sigma\in H(\mathrm{div};\Omega)$ et $z^u\in H^1(\Omega)$.

Il est clair que l'espace W muni de la norme (et du produit scalaire associé)

$$\|z\|_W^2=\|z^\sigma\|_{L^2}^2+\|\nabla\cdot z^\sigma\|_{L^2}^2+\|z^u\|_{H^1}^2 \tag{56}$$

est un espace de Hilbert. De plus, on dispose du théorème de trace suivant : l'application

$$\gamma: W\ni z\longmapsto (n\cdot z^\sigma,z^u)\in H^{-\frac12}(\Omega)\times H^{\frac12}(\Omega) \tag{57}$$

est continue sur W. Enfin, on définit sur $W\times W$ la forme bilinéaire

$$a(z,y)=\int_\Omega(z^\sigma+\nabla z^u)\cdot y^\sigma+\int_\Omega(\nabla\cdot z^\sigma+z^u)y^u-\langle n\cdot y^\sigma,z^u\rangle_{H^{-\frac12},H^{\frac12}}, \tag{58}$$

qui est bien continue sur $W\times W$. De plus, cette forme linéaire est L^2-coercive sur W puisqu'en utilisant la formule de Green, il vient

$$a(z,z)=\|z^\sigma\|_{L^2}^2+\|z^u\|_{L^2}^2. \tag{59}$$

On considère le problème suivant :

$$\text{Chercher } z\in W \text{ tel que } a(z,y)=\int_\Omega f y^u,\ \forall y\in W. \tag{60}$$

Nous avons le résultat suivant (la preuve est laissée en exercice ; on utilisera la surjectivité des traces de $H(\mathrm{div};\Omega)$ pour récupérer la conditions aux limites).

Théorème 8. *Le problème* (60) *est bien posé. De plus, son unique solution est solution de* (54).

2.2 Construction de l'approximation GD

On pose $U_h=\mathbb{G}_h^p$, $\Sigma_h=[\mathbb{G}_h^p]^d$ et $W_h=\Sigma_h\times U_h$. Le problème discret consiste à

$$\text{Chercher } z_h\in W_h \text{ tel que } a_h(z_h,y_h)=\int_\Omega f y_h^u,\ \forall y_h\in W_h. \tag{61}$$

On considère donc une approximation de GD de degré p pour la variable primale et pour le flux. Deux hypothèses de régularité supplémentaires sont faites sur la solution exacte :

- $z:=(\sigma,u)\in[H^{p+1}(\mathcal{T}_h)]^d$, ce qui signifie que le maillage a été construit de sorte que la solution exacte soit suffisamment régulière à l'intérieur de chaque maille ;
- pour tout $F\in\mathcal{F}_h^{\mathrm{i}}$, $n_F\cdot[\![\sigma]\!]=0$ et $[\![u]\!]=0$.

Examinons la L^2-coercivité de la forme bilinéaire a définie par (58) sur W_h en prenant soin, au préalable, de localiser les opérateurs de gradient et de divergence sur chaque maille. On a pour tout $z_h = (z_h^\sigma, z_h^u) \in W_h$,

$$\begin{aligned} a(z_h, z_h) &= \|z_h^\sigma\|_{L^2}^2 + \|z_h^u\|_{L^2}^2 + \sum_{F\in\mathcal{F}_h^{\rm i}} \int_F 2\{n\cdot z_h^\sigma z_h^u\} \\ &= \|z_h^\sigma\|_{L^2}^2 + \|z_h^u\|_{L^2}^2 + \sum_{F\in\mathcal{F}_h^{\rm i}} \int_F n_F\cdot[\![z_h^\sigma]\!]\{z_h^u\} + n_F\cdot\{z_h^\sigma\}[\![z_h^u]\!]. \end{aligned}$$

Puisque la solution exacte satisfait $n_F\cdot[\![\sigma]\!] = 0$ et $[\![u]\!] = 0$ pour tout $F \in \mathcal{F}_h^{\rm i}$, on rétablit la L^2-coercivité sur W_h tout en préservant la consistance en considérant la forme bilinéaire

$$\begin{aligned} a_h(z_h, y_h) = &\int_\Omega (z_h^\sigma + \nabla_h z_h^u)\cdot y_h^\sigma + \int_\Omega (\nabla_h\cdot z_h^\sigma + z_h^u) y_h^u - \sum_{F\in\mathcal{F}_h^\partial} \int_F (n\cdot y_h^\sigma) z_h^u \\ &- \sum_{F\in\mathcal{F}_h^{\rm i}} \int_F n_F\cdot[\![z_h^\sigma]\!]\{y_h^u\} + n_F\cdot\{y_h^\sigma\}[\![z_h^u]\!]. \end{aligned} \tag{62}$$

(Observer les choix faits pour z_h et y_h dans les termes relatifs aux interfaces.) On a donc

$$a_h(z_h, z_h) \geq \|z_h^\sigma\|_{L^2}^2 + \|z_h^u\|_{L^2}^2, \qquad \forall z_h \in W_h, \tag{63}$$

$$a_h(z, y_h) = \int_\Omega f y_h^u, \qquad \forall y_h \in W_h. \tag{64}$$

Il reste à pénaliser au sens des moindres carrés les sauts aux interfaces de z_h^u et de la composante normale de z_h^σ. De plus, comme l'inégalité (63) ne fournit aucun contrôle sur les valeurs au bord de z_h^u, ces dernières seront également pénalisées au sens des moindres carrés. Etant donné deux paramètres $\eta^u > 0$ et $\eta^\sigma > 0$ choisis par le numéricien (et qui peuvent éventuellement être pris différents sur chaque face), on considère la forme bilinéaire

$$\begin{aligned} a_h(z_h, y_h) = &\int_\Omega (z_h^\sigma + \nabla_h z_h^u)\cdot y_h^\sigma + \int_\Omega (\nabla_h\cdot z_h^\sigma + z_h^u) y_h^u - \sum_{F\in\mathcal{F}_h^\partial} \int_F (n\cdot y_h^\sigma) z_h^u \\ &- \sum_{F\in\mathcal{F}_h^{\rm i}} \int_F n_F\cdot[\![z_h^\sigma]\!]\{y_h^u\} + n_F\cdot\{y_h^\sigma\}[\![z_h^u]\!] \\ &+ \sum_{F\in\mathcal{F}_h} \int_F \eta^u [\![z_h^u]\!][\![y_h^u]\!] + \sum_{F\in\mathcal{F}_h^{\rm i}} \int_F \eta^\sigma (n_F\cdot[\![z_h^\sigma]\!])(n_F\cdot[\![y_h^\sigma]\!]), \end{aligned} \tag{65}$$

ou encore en intégrant par parties le gradient et la divergence sur chaque maille,

$$\begin{aligned} a_h(z_h, y_h) = &\int_\Omega [z_h^\sigma\cdot y_h^\sigma - z_h^u \nabla_h\cdot y_h^\sigma] + \int_\Omega [z_h^u y_h^u - z_h^\sigma\cdot\nabla_h y_h^u] + \sum_{F\in\mathcal{F}_h^\partial} \int_F (n\cdot z_h^\sigma) y_h^u \\ &+ \sum_{F\in\mathcal{F}_h^{\rm i}} \int_F n_F\cdot[\![y_h^\sigma]\!]\{z_h^u\} + n_F\cdot\{z_h^\sigma\}[\![y_h^u]\!] \\ &+ \sum_{F\in\mathcal{F}_h} \int_F \eta^u [\![z_h^u]\!][\![y_h^u]\!] + \sum_{F\in\mathcal{F}_h^{\rm i}} \int_F \eta^\sigma (n_F\cdot[\![z_h^\sigma]\!])(n_F\cdot[\![y_h^\sigma]\!]). \end{aligned} \tag{66}$$

Avant de passer à l'analyse de convergence, examinons la formulation locale du problème (61) et les flux associés. On vérifie facilement en utilisant (66) que $z_h := (z_h^\sigma, z_h^u)$ est solution de (61) si et seulement si pour toute maille $T \in \mathcal{T}_h$, pour tout polynôme $q \in [\mathbb{P}_p]^d$ et pour tout polynôme $r \in \mathbb{P}_p$,

$$\begin{aligned} \int_T [z_h^\sigma \cdot q - z_h^u \nabla \cdot q] + \sum_{F \subset \partial T} \int_F \phi_{T,F}^\sigma(z_h) \cdot q = 0, \\ \int_T [z_h^u r - z_h^\sigma \cdot \nabla r] + \sum_{F \subset \partial T} \int_F \phi_{T,F}^u(z_h) r = \int_T f r, \end{aligned} \tag{67}$$

avec les flux

$$\phi_{T,F}^\sigma(z_h) = \begin{cases} n_T(\{z_h^u\} + \eta^\sigma n_F \cdot [\![z_h^\sigma]\!]), & F \in \mathcal{F}_h^{\mathrm{i}}, \\ 0, & F \in \mathcal{F}_h^\partial, \end{cases} \tag{68}$$

et

$$\phi_{T,F}^u(z_h) = \begin{cases} n_T \cdot (\{z_h^\sigma\} + \eta^u n_F [\![z_h^u]\!]), & F \in \mathcal{F}_h^{\mathrm{i}}, \\ n \cdot (z_h^\sigma + \eta^u n z_h^u), & F \in \mathcal{F}_h^\partial. \end{cases} \tag{69}$$

Ces flux sont bien conservatifs puisque si $F = T_1 \cap T_2$,

$$\phi_{T_1,F}^\sigma(z_h) + \phi_{T_2,F}^\sigma(z_h) = 0, \qquad \phi_{T_1,F}^u(z_h) + \phi_{T_2,F}^u(z_h) = 0. \tag{70}$$

Ils sont de plus consistants avec les fluxs exacts qui sont obtenus formellement en intégrant par parties sur une maille $T \in \mathcal{T}_h$ les EDPs $\sigma + \nabla u = 0$ et $\nabla \cdot \sigma + u = f$ multipliées respectivement par un polynôme $q \in [\mathbb{P}_p]^d$ et un polynôme $r \in \mathbb{P}_p$, ce qui donne

$$\begin{aligned} \int_T [z^\sigma \cdot q - z^u \nabla \cdot q] + \sum_{F \subset \partial T} \int_F \Phi_{T,F}^\sigma(z) \cdot q = 0, \\ \int_T [z^u r - z^\sigma \cdot \nabla r] + \sum_{F \subset \partial T} \int_F \Phi_{T,F}^u(z) r = \int_T f r, \end{aligned} \tag{71}$$

avec

$$\Phi_{T,F}^\sigma(z) = n_T z^u, \qquad \Phi_{T,F}^u(z) = n_T \cdot z^\sigma. \tag{72}$$

Grâce aux hypothèses de régularité sur la solution exacte et aux conditions aux limites, on a bien

$$\phi_{T,F}^\sigma(z) = \Phi_{T,F}^\sigma(z), \qquad \phi_{T,F}^u(z) = \Phi_{T,F}^u(z). \tag{73}$$

2.3 Analyse de convergence

La norme de stabilité est la suivante :

$$\begin{aligned} |\!|\!| z |\!|\!|^2 = \|z^\sigma\|_{L^2}^2 + \|z^u\|_{L^2}^2 + |z^u|_M^2 + |z^u|_{J^u}^2 + |z^\sigma|_{J^\sigma}^2 \\ + \sum_{T \in \mathcal{T}_h} [h_T \|\nabla z^u\|_T^2 + h_T \|\nabla \cdot z^\sigma\|_T^2], \end{aligned} \tag{74}$$

avec

$$|z^u|_M^2 = \sum_{F\in\mathcal{F}_h^\partial} |z^u|_{M,F}^2, \qquad |z^u|_{M,F}^2 = \int_F \eta^u (z^u)^2, \tag{75}$$

$$|z^u|_{J^u}^2 = \sum_{F\in\mathcal{F}_h^{\rm i}} |z^u|_{J^u,F}^2, \qquad |z^u|_{J^u,F}^2 = \int_F \eta^u [\![z^u]\!]^2, \tag{76}$$

$$|z^\sigma|_{J^\sigma}^2 = \sum_{F\in\mathcal{F}_h^{\rm i}} |z^\sigma|_{J^\sigma,F}^2, \qquad |z^\sigma|_{J^\sigma,F}^2 = \int_F \eta^\sigma (n_F\cdot[\![z^\sigma]\!])^2. \tag{77}$$

Lemme 9 (Stabilité)**.** *On a*

$$\forall z_h \in W_h, \qquad |\!|\!| z_h |\!|\!| \lesssim \sup_{y_h\in W_h\setminus\{0\}} \frac{a_h(z_h,y_h)}{|\!|\!| y_h |\!|\!|}. \tag{78}$$

Preuve. Soit $z_h \in W_h$. On pose $\mathbb{S} = \sup_{y_h\in W_h\setminus\{0\}} \frac{a_h(z_h,y_h)}{|\!|\!| y_h |\!|\!|}$.
(i) Il est clair que

$$\|z_h^\sigma\|_{L^2}^2 + \|z_h^u\|_{L^2}^2 + |z_h^u|_M^2 + |z_h^u|_{J^u}^2 + |z_h^\sigma|_{J^\sigma}^2 \lesssim a_h(z_h,z_h) \lesssim \mathbb{S}|\!|\!| z_h |\!|\!|.$$

(ii) Il reste à contrôler la norme du gradient de z_h^u et celle de la divergence de z_h^σ. Pour cela, on procède de manière analogue à la preuve de la condition inf-sup discrète pour le problème d'advection–réaction. Pour obtenir le contrôle sur la divergence de z_h^σ, on introduit le champ auxiliaire π_h tel que sur chaque maille $T \in \mathcal{T}_h$,

$$\pi_h|_T = (\pi_h^\sigma, \pi_h^u)|_T := (0, h_T\nabla\cdot z_h^\sigma)|_T.$$

Il est clair que $\pi_h \in W_h$. De plus, en utilisant des inégalités inverse et de trace, on vérifie que $|\!|\!| \pi_h |\!|\!| \lesssim |\!|\!| z_h |\!|\!|$. Enfin, on observe que

$$\sum_{T\in\mathcal{T}_h} h_T\|\nabla\cdot z_h^\sigma\|_T^2 = a_h(z_h,\pi_h) - (z_h^u,\pi_h^u)_\Omega + \sum_{F\in\mathcal{F}_h^{\rm i}} \int_F n_F\cdot[\![z_h^\sigma]\!]\{\pi_h^u\}$$
$$- \sum_{F\in\mathcal{F}_h} \int_F \eta^u [\![z_h^u]\!][\![\pi_h^u]\!],$$

et en majorant les quatre termes du membre de droite, on obtient (les détails sont laissés en exercice),

$$\sum_{T\in\mathcal{T}_h} h_T\|\nabla\cdot z_h^\sigma\|_T^2 \lesssim \mathbb{S}^{\frac12}|\!|\!| z_h |\!|\!|^{\frac32} + \mathbb{S}|\!|\!| z_h |\!|\!|.$$

Par ailleurs, on obtient un contrôle analogue sur le gradient de z_h^u en considérant le champ auxiliaire π_h tel que sur chaque maille $T \in \mathcal{T}_h$,

$$\pi_h|_T = (\pi_h^\sigma, \pi_h^u)|_T := (h_T\nabla z_h^u, 0)|_T,$$

ce qui conduit à

$$\sum_{T\in\mathcal{T}_h} h_T\|\nabla z_h^u\|_T^2 \lesssim \mathbb{S}^{\frac12}|\!|\!| z_h |\!|\!|^{\frac32} + \mathbb{S}|\!|\!| z_h |\!|\!|.$$

(iii) En collectant les majorations ci-dessus, il vient

$$|||z_h|||^2 \lesssim \mathbb{S}^{\frac{1}{2}}|||z_h|||^{\frac{3}{2}} + \mathbb{S}|||z_h|||.$$

et on conclut en utilisant deux fois l'inégalité de Young. □

Pour la propriété de continuité, on utilise le norme

$$|||z|||_*^2 = |||z|||^2 + \sum_{T\in\mathcal{T}_h} [h_T^{-1}\|z^u\|_T^2 + \|z^u\|_{\partial T}^2 + h_T^{-1}\|z^\sigma\|_T^2 + \|z^\sigma\|_{\partial T}^2]. \tag{79}$$

On rappelle que l'espace Z où vit l'erreur est $Z = [H^{p+1}(\mathcal{T}_h)]^{d+1} + W_h$.

Lemme 10 (Continuité). *On a*

$$\forall \zeta \in Z,\ \forall y_h \in W_h, \qquad a_h(\zeta, y_h) \lesssim |||\zeta|||_* |||y_h|||. \tag{80}$$

Preuve. Vérification directe en majorant un à un tous les termes qui composent le membre de droite de (66). Les détails sont laissés en exercice. □

Nous sommes maintenant en mesure d'énoncer le résultat principal de cette section.

Théorème 11 (Convergence). *On a*

$$|||z - z_h||| \lesssim h^{p+\frac{1}{2}}\|z\|_{[H^{p+1}(\mathcal{T}_h)]^{d+1}}. \tag{81}$$

Preuve. Utiliser le cadre théorique ci-dessus pour obtenir

$$|||z - z_h||| \lesssim \inf_{y_h\in W_h} |||z - y_h|||_*,$$

puis utiliser le résultat d'interpolation (16) pour estimer le membre de droite. □

Le tableau 2.3 résume les ordres de convergence obtenus pour les erreurs en flux et en variable primale. On notera que les ordres de convergence sont légèrement suboptimaux (d'ordre $\frac{1}{2}$) pour la norme L^2 et qu'ils sont optimaux aussi bien pour le gradient de la variable primale que pour la divergence du flux.

$\|z^\sigma - z_h^\sigma\|_\Omega$	$\|\nabla_h\cdot(z^\sigma - z_h^\sigma)\|_\Omega$	$\|z^u - z_h^u\|_\Omega$	$\|\nabla_h(z^u - z_h^u)\|_\Omega$
$\mathcal{O}(h^{p+\frac{1}{2}})$	$\mathcal{O}(h^p)$	$\mathcal{O}(h^{p+\frac{1}{2}})$	$\mathcal{O}(h^p)$

TAB. 1 – Ordres de convergence pour l'approximation GD à deux champs ; pénalisation de $[\![z_h^u]\!]$ avec un coefficient $\eta^u > 0$ et pénalisation de $n_F\cdot[\![z_h^\sigma]\!]$ avec un coefficient $\eta^\sigma > 0$; la solution exacte est supposée être dans $[H^{p+1}(\mathcal{T}_h)]^d \times H^{p+1}(\mathcal{T}_h)$.

3 Diffusion–réaction : approche à un champ

L'approximation GD à deux champs présentée dans la section précédente pour le problème de diffusion–réaction présente l'inconvénient d'être onéreuse. En effet, la taille du système linéaire à résoudre est de $(d+1)C_{p+d}^d N_{\mathrm{ma}}$. Par exemple, pour une approximation de degré un, on obtient une taille de $9N_{\mathrm{ma}}$ en dimension 2 et de $16N_{\mathrm{ma}}$ en dimension 3, ce qui est énorme.

L'approche développée dans cette section permet de pallier à cette difficulté. Elle consiste à réduire considérablement la taille du système linéaire en éliminant le flux discret localement sur chaque maille. Le prix à payer est que cette élimination est rendue possible

- en ne pénalisant plus le saut de la composante normale du flux aux interfaces ;
- en pénalisant le saut de la variable primale par un coefficient qui explose en $\frac{1}{h}$.

Sur le plan de l'analyse d'erreur, la conséquence de cette modification dans la stratégie de pénalisation est que l'ordre de convergence pour le flux sera dégradé, l'ordre de convergence pour la variable primale restant identique à celui fourni par l'approximation GD à deux champs.

3.1 Elimination locale du flux

Afin de bien comprendre comment est rendue possible l'élimination locale du flux discret z_h^σ, considérons la première équation dans la formulation locale (67) (celle faisant intervenir le polynôme q). Celle-ci permet d'exprimer $z_h^\sigma|_T$ uniquement en fonction de la variable primale si le flux $\phi_{T,F}^\sigma$ ne dépend pas des valeurs prises par z_h^σ dans les mailles voisines. En examinant (68), on voit que cela est possible en choisissant

$$\eta^\sigma = 0. \tag{82}$$

Dans ces conditions, on constate en évaluant le membre de droite de (65) pour $y_h = (y_h^\sigma, 0)$, qu'on obtient

$$\int_\Omega (z_h^\sigma + \nabla_h z_h^u)\cdot y_h^\sigma - \sum_{F\in\mathcal{F}_h^\partial}\int_F (n\cdot y_h^\sigma) z_h^u - \sum_{F\in\mathcal{F}_h^{\mathrm{i}}}\int_F n_F\cdot\{y_h^\sigma\}[\![z_h^u]\!] = 0. \tag{83}$$

Soit $F\in\mathcal{F}_h$. On introduit l'opérateur de relèvement

$$r_F : [L^2(F)]^d \ni v \longmapsto r_F(v)\in\Sigma_h, \tag{84}$$

où $r_F(v)$ est tel que

$$\forall \tau_h\in\Sigma_h, \qquad \int_\Omega r_F(v)\cdot\tau_h = \int_F v\cdot\{\tau_h\}. \tag{85}$$

(On rappelle que pour $F\in\mathcal{F}_h^\partial$, on a par convention $\{\tau_h\}=\tau_h$.) On notera que le support de r_F est Δ_F, c'est-à-dire les deux mailles partageant F pour $F\in\mathcal{F}_h^{\mathrm{i}}$ et la seule maille dont F est une face pour $F\in\mathcal{F}_h^\partial$. Puis, on introduit l'opérateur

$$R_h : U_h \ni u_h \longmapsto R_h(u_h) = \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} r_F(n_F[\![u_h]\!]) + \sum_{F\in\mathcal{F}_h^\partial} r_F(nu_h). \tag{86}$$

On notera que si le support de u_h est inclus dans une maille $T\in\mathcal{T}_h$, le support de $R_h(u_h)$ est constitué de la réunion de T et des mailles de $\mathcal{T}_h$ partageant une face avec T.

L'intérêt d'introduire l'opérateur R_h est qu'il permet de réécrire l'équation (83) sous la forme compacte

$$z_h^\sigma = -\nabla_h z_h^u + R_h(z_h^u). \tag{87}$$

On constate que le flux discret est égal au gradient local de la variable primale corrigé par un relèvement des sauts de la variable primale sur les faces.

Evaluons maintenant le membre de droite de (66) pour $y_h = (0, y_h^u)$ en y substituant l'expression de z_h^σ donnée par (87). Il vient

$$\begin{aligned}
a_h(z_h,(0,y_h^u)) = &\int_\Omega [z_h^u y_h^u - z_h^\sigma{\cdot}\nabla_h y_h^u] + \sum_{F\in\mathcal{F}_h^\partial}\int_F (n{\cdot}z_h^\sigma) y_h^u \\
&+ \sum_{F\in\mathcal{F}_h^{\rm i}}\int_F n_F{\cdot}\{z_h^\sigma\}[\![y_h^u]\!] + \sum_{F\in\mathcal{F}_h}\int_F \eta^u [\![z_h^u]\!][\![y_h^u]\!] \\
= &\int_\Omega [z_h^u y_h^u + (\nabla_h z_h^u - R_h(z_h^u)){\cdot}\nabla_h y_h^u] + \sum_{F\in\mathcal{F}_h^\partial}\int_F (n{\cdot}z_h^\sigma) y_h^u \\
&+ \sum_{F\in\mathcal{F}_h^{\rm i}}\int_F n_F{\cdot}\{z_h^\sigma\}[\![y_h^u]\!] + \sum_{F\in\mathcal{F}_h}\int_F \eta^u [\![z_h^u]\!][\![y_h^u]\!].
\end{aligned}$$

Or,

$$\begin{aligned}
\sum_{F\in\mathcal{F}_h^\partial}\int_F (n{\cdot}z_h^\sigma) y_h^u + \sum_{F\in\mathcal{F}_h^{\rm i}}\int_F n_F{\cdot}\{z_h^\sigma\}[\![y_h^u]\!] &= \sum_{F\in\mathcal{F}_h^\partial}\int_\Omega r_F(n y_h^u){\cdot}z_h^\sigma + \sum_{F\in\mathcal{F}_h^{\rm i}}\int_\Omega r_F(n_F[\![y_h^u]\!]){\cdot}z_h^\sigma \\
&= \int_\Omega R_F(y_h^u){\cdot}z_h^\sigma.
\end{aligned}$$

Par suite, si on pose

$$a_h^{\rm LDG}(z_h^u, y_h^u) = \int_\Omega [z_h^u y_h^u + (\nabla_h z_h^u - R_h(z_h^u)){\cdot}(\nabla_h y_h^u - R_h(y_h^u))] + \sum_{F\in\mathcal{F}_h}\int_F \eta^u [\![z_h^u]\!][\![y_h^u]\!], \tag{88}$$

on a montré le résultat suivant.

Proposition 12. *On suppose que $\eta^\sigma = 0$. Alors, $z_h = (z_h^\sigma, z_h^u)$ est solution du problème discret* (61) *si et seulement si z_h^u est solution du problème qui consiste à*

$$\text{Chercher } z_h^u \in U_h \text{ tel que } a_h^{\rm LDG}(z_h^u, y_h^u) = \int_\Omega f y_h^u,\ \forall y_h^u \in U_h, \tag{89}$$

et si z_h^σ est évalué selon (87).

La méthode de GD ainsi construite porte le nom de LDG de l'acronyme anglais «Local Discontinuous Galerkin». Le système linéaire associé à (89) est de taille $C_{p+d}^d N_{\rm ma}$, soit un gain d'un facteur $(d+1)$ par rapport au problème (61).

3.2 Analyse de convergence

Afin d'obtenir des estimations d'erreur satisfaisantes pour la méthode LDG, le coefficient de pénalisation η^u doit être choisi de sorte que

$$\forall F \in \mathcal{F}_h, \qquad \eta^u = \frac{\alpha}{h_F}, \tag{90}$$

où $\alpha > 0$ est un coefficient choisi par le numéricien et h_F le diamètre de la face F.

Pour la méthode LDG, les normes de stabilité et de continuité sont les suivantes :

$$\begin{aligned} |||z|||^2 &= \|z^\sigma\|_{L^2}^2 + \|z^u\|_{L^2}^2 + |z^u|_M^2 + |z^u|_{J^u}^2 + \sum_{T\in\mathcal{T}_h} \|\nabla z^u\|_T^2, \\ |||z|||_*^2 &= |||z|||^2 + \sum_{T\in\mathcal{T}_h} [h_T^{-2}\|z^u\|_T^2 + h_T^{-1}\|z^u\|_{\partial T}^2 + h_T\|z^\sigma\|_{\partial T}^2], \end{aligned} \tag{91}$$

avec

$$|z^u|_M^2 = \sum_{F\in\mathcal{F}_h^\partial} |z^u|_{M,F}^2, \qquad |z^u|_{M,F}^2 = \int_F \tfrac{\alpha}{h_F}(z^u)^2, \tag{92}$$

$$|z^u|_{J^u}^2 = \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} |z^u|_{J^u,F}^2, \qquad |z^u|_{J^u,F}^2 = \int_F \tfrac{\alpha}{h_F}[\![z^u]\!]^2. \tag{93}$$

La preuve des résultats ci-dessous est laissée en exercice. Elle se fonde sur les mêmes techniques que précédemment. La seule différence est que l'hypothèse de régularité sur la solution exacte peut être affaiblie à $z \in [H^p(\mathcal{T}_h)]^d \times H^{p+1}(\mathcal{T}_h)$.

Lemme 13 (Stabilité). *On a*

$$\forall z_h \in W_h, \qquad |||z_h||| \lesssim \sup_{y_h\in W_h\setminus\{0\}} \frac{a_h^{\mathrm{LDG}}(z_h, y_h)}{|||y_h|||}. \tag{94}$$

Lemme 14 (Continuité). *On a*

$$\forall \zeta \in Z,\ \forall y_h \in W_h, \qquad a_h^{\mathrm{LDG}}(\zeta, y_h) \lesssim |||\zeta|||_* |||y_h|||. \tag{95}$$

Théorème 15 (Convergence). *On a*

$$|||z - z_h||| \lesssim h^p \|z\|_{[H^p(\mathcal{T}_h)]^d\times H^{p+1}(\mathcal{T}_h)}. \tag{96}$$

Le tableau 3.2 résume les ordres de convergence obtenus pour les erreurs en flux et en variable primale (comparer avec le tableau 2.3). On notera que les ordres de convergence sont suboptimaux (d'ordre 1) pour la norme L^2 et qu'ils sont optimaux pour le gradient de la variable primale. Lorsque $p \geq 2$, une estimation d'erreur d'ordre h^{p-1} peut être établie pour la divergence du flux en utilisant une inégalité inverse et l'estimation de $\|z^\sigma - z_h^\sigma\|_\Omega$. Enfin, lorsqu'on dispose d'une propriété de régularité elliptique, on peut améliorer l'estimation de l'erreur $\|z^u - z_h^u\|_\Omega$ par des techniques de dualité et obtenir pour celle-ci une convergence optimale d'ordre $(p+1)$.

Remarque 16. *Dans la méthode LDG, il est possible de choisir $\Sigma_h = [\mathbb{G}_h^{p-1}]^d$, c'est-à-dire une approximation polynomiale d'un degré inférieure pour le flux par rapport à celle de la variable primale. Les estimations d'erreur ci-dessus ne s'en trouvent pas modifiées. L'avantage est que l'estimation de $\|z^\sigma - z_h^\sigma\|_\Omega$ devient du coup optimale. Toutefois, une méthode LDG utilsant des degrés polynomiaux différents pour le flux et la variable primale peut être moins simple à programmer.*

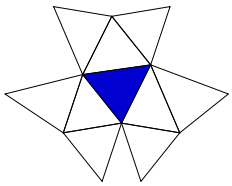
$\|z^\sigma - z_h^\sigma\|_\Omega$	$\|z^u - z_h^u\|_\Omega$	$\|\nabla_h(z^u - z_h^u)\|_\Omega$
$\mathcal{O}(h^p)$	$\mathcal{O}(h^p)$	$\mathcal{O}(h^p)$

TAB. 2 – Ordres de convergence pour l'approximation GD à un champ ; pénalisation de $[\![z_h^u]\!]$ avec un coefficient $\eta^u = \mathcal{O}(\frac{1}{h})$ et pas de pénalisation de $n_F{\cdot}[\![z_h^\sigma]\!]$; la solution exacte est supposée être dans $[H^p(\mathcal{T}_h)]^d \times H^{p+1}(\mathcal{T}_h)$.

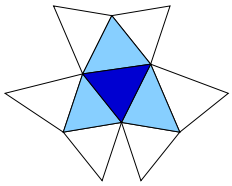
3.3 Réduction du stencil

Un désavantage de la méthode LDG est que dans la matrice du système linéaire associé, les degrés de liberté sur une maille sont couplés à ceux des mailles voisines mais également à ceux des mailles voisines de ces dernières. Considérons l'expression (88) pour a_h^{LDG} et examinons le stencil associé à chacun des termes après avoir développé le terme contenant les gradients locaux et les relèvements. On illustre ci-dessus pour chaque terme le stencil associé ; celui-ci s'obtient en considérant une fonction test y_h^u dont le support est localisé sur la maille au centre du macro-élément et en cherchant tous les triangles pour lesquels les degrés de liberté locaux apportent une contribution non-nulle ; ceux-ci sont représentés en teinte grisée.

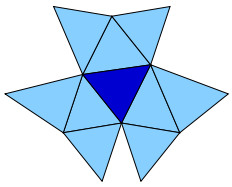
$$\int_\Omega [z_h^u y_h^u + \nabla_h z_h^u{\cdot}\nabla_h y_h^u]$$



$$\int_\Omega [R_h(z_h^u){\cdot}\nabla_h y_h^u + \nabla_h z_h^u{\cdot}R_h(y_h^u)] + \sum_{F\in\mathcal{F}_h} \int_F \tfrac{\alpha}{h_F}[\![z_h^u]\!][\![y_h^u]\!]$$



$$\int_\Omega R_h(z_h^u){\cdot}R_h(y_h^u)$$



On constate donc que le seul terme responsable de l'extension du stencil est le terme

$$\rho_h(z_h^u, y_h^u) := \int_\Omega R_h(z_h^u){\cdot}R_h(y_h^u). \tag{97}$$

Celui-ci fait passer le stencil de 4 à 10 triangles en dimension 2 et de 5 à 13 tétraèdres en dimension 3. D'où l'idée d'éliminer ce terme dans la forme bilinéaire. On pose

$$a_h^{\mathrm{SIPG}}(z_h^u, y_h^u) = a_h^{\mathrm{LDG}}(z_h^u, y_h^u) - \rho_h(z_h^u, y_h^u). \tag{98}$$

En observant que

$$\int_\Omega R_h(z_h^u){\cdot}\nabla_h y_h^u = \sum_{F\in\mathcal{F}_h} \int_F n_F{\cdot}\{\nabla_h y_h^u\}[\![z_h^u]\!],$$

il vient

$$a_h^{\mathrm{SIPG}}(z_h^u, y_h^u) = \int_\Omega [z_h^u y_h^u + \nabla_h z_h^u \cdot \nabla_h y_h^u] + \sum_{F\in\mathcal{F}_h} \int_F \tfrac{\alpha}{h_F} [\![z_h^u]\!][\![y_h^u]\!] \\ - \sum_{F\in\mathcal{F}_h} \int_F [n_F \cdot \{\nabla_h y_h^u\}[\![z_h^u]\!] + n_F \cdot \{\nabla_h z_h^u\}[\![y_h^u]\!]]. \tag{99}$$

On reconnaît la forme bilinéaire associée à la méthode SIPG, de l'acronyme anglais «Symmetric Interior Penalty Galerkin».

On considère le problème discret

$$\text{Chercher } z_h^u \in U_h \text{ tel que } a_h^{\mathrm{SIPG}}(z_h^u, y_h^u) = \int_\Omega f y_h^u,\ \forall y_h^u \in U_h. \tag{100}$$

L'analyse de convergence de ce schéma nécessite, comme toujours, d'établir des propriétés de stabilité et de continuité pour la forme bilinéaire a_h^{SIPG}. Pour cela, on utilise le lien entre les formes bilinéaires a_h^{SIPG} et a_h^{LDG} (voir (98)) et les résultats des lemmes 13 et 14 pour la forme bilinéaire a_h^{LDG}.

Lemme 17. *On désigne par N_∂ le nombre de faces d'une maille $T \in \mathcal{T}_h$. On a pour tout $u_h \in U_h$,*

$$\|R_h(u_h)\|_\Omega^2 \le N_\partial \sum_{F\in\mathcal{F}_h} \|r_F(n_F[\![u_h]\!])\|_\Omega^2. \tag{101}$$

Preuve. Observer que

$$\begin{aligned} \|R_h(u_h)\|_\Omega^2 &= \sum_{T\in\mathcal{T}_h} \int_T \left(\sum_{F\subset\partial T} r_F(n_F[\![u_h]\!]) \right)^2 \\ &\le \sum_{T\in\mathcal{T}_h} \int_T N_\partial \sum_{F\subset\partial T} |r_F(n_F[\![u_h]\!])|^2 \\ &= N_\partial \sum_{F\in\mathcal{F}_h} \|r_F(n_F[\![u_h]\!])\|_\Omega^2. \end{aligned}$$

□

Lemme 18. *On a pour tout $u_h \in U_h$,*

$$\sum_{F\in\mathcal{F}_h} \|r_F(n_F[\![u_h]\!])\|_\Omega^2 \lesssim |u_h|_M^2 + |u_h|_{J^u}^2. \tag{102}$$

Preuve. Observer que par définition, pour tout $F \in \mathcal{F}_h$ et pour tout $v \in [L^2(F)]^d$,

$$\int_\Omega |r_F(v)|^2 = \int_F v\{r_F(v)\},$$

puis utiliser une inégalité de trace pour conclure que $\|r_F(n_F[\![u_h]\!])\|_\Omega \lesssim |u_h|_{J^u,F}$ si $F \in \mathcal{F}_h^{\mathrm{i}}$ et $\|r_F(n_F[\![u_h]\!])\|_\Omega \lesssim |u_h|_{M,F}$ si $F \in \mathcal{F}_h^\partial$. □

Il en résulte le résultat de convergence suivant.

Théorème 19. *Il existe $\alpha_0 > 0$ dépendant de la valeur de la constante intervenant dans l'inégalité de trace* (17) *telle que si $\alpha \geq \alpha_0$, le problème discret* (100) *est bien posé et son unique solution satisfait les conclusions du théorème* 15.

Preuve. Grâce aux lemmes 17 et 18, on vérifie que sous l'hypothèse $\alpha \geq \alpha_0$, la forme bilinéaire a_h^{SIPG} vérifie la propriété de L^2-coercivité suivante : pour tout $u_h \in U_h$,

$$a_h^{\mathrm{SIPG}}(u_h, u_h) \gtrsim \|\nabla_h u_h\|_{L^2}^2 + \|u_h\|_{L^2}^2 + |u_h|_M^2 + |u_h|_{J^u}^2.$$

On conclut aisément. □

On retiendra que la méthode SIPG permet une réduction substantielle du coût des calculs par rapport à la méthode LDG en réduisant la taille du stencil tout en conduisant aux mêmes estimations d'erreur ; le prix à payer est que le coefficient intervenant dans la pénalisation des sauts de la variable primale doit être suffisamment grand, sa valeur minimale dépendant de la valeur de la constante intervenant dans l'inégalité de trace (17). Toutefois, il est possible de s'affranchir de cette dépendance en considérant une variante de la méthode SIPG, dite méthode BRMPS (du nom de ses auteurs, voir la bibliographie). On pose

$$\begin{aligned} a_h^{\mathrm{BRMPS}}(z_h^u, y_h^u) = &\int_\Omega [z_h^u y_h^u + \nabla_h z_h^u \cdot \nabla_h y_h^u] + \sum_{F \in \mathcal{F}_h} \int_\Omega \alpha r_F(n_F[\![z_h^u]\!]) \cdot r_F(n_F[\![y_h^u]\!]) \\ &- \sum_{F \in \mathcal{F}_h} \int_F [n_F \cdot \{\nabla_h y_h^u\}[\![z_h^u]\!] + n_F \cdot \{\nabla_h z_h^u\}[\![y_h^u]\!]], \end{aligned} \tag{103}$$

et on considère le problème discret

$$\text{Chercher } z_h^u \in U_h \text{ tel que } a_h^{\mathrm{BRMPS}}(z_h^u, y_h^u) = \int_\Omega f y_h^u, \ \forall y_h^u \in U_h. \tag{104}$$

En utilisant le lemme 17, on vérifie sans peine le résultat suivant.

Théorème 20. *Sous l'hypothèse $\alpha > N_\partial$, le problème discret* (104) *est bien posé et son unique solution satisfait les conclusions du théorème* 15.

Concluons cette section par une comparaison des flux intervenant dans les problèmes locaux associés aux méthodes LDG et SIPG. Pour la méthode LDG, les flux se déduisent de (68) et (69) en tenant compte du fait que $\eta^\sigma = 0$ et $\eta^u = \frac{\alpha}{h}$. Il vient

$$\phi_{T,F}^\sigma(z_h) = \begin{cases} n_T\{z_h^u\}, & F \in \mathcal{F}_h^{\mathrm{i}}, \\ 0, & F \in \mathcal{F}_h^\partial, \end{cases} \tag{105}$$

et

$$\phi_{T,F}^u(z_h) = \begin{cases} n_T \cdot (\{z_h^\sigma\} + \frac{\alpha}{h} n_F[\![z_h^u]\!]), & F \in \mathcal{F}_h^{\mathrm{i}}, \\ n \cdot (z_h^\sigma + \frac{\alpha}{h} n z_h^u), & F \in \mathcal{F}_h^\partial. \end{cases} \tag{106}$$

Pour la méthode SIPG, on observe pour une maille $T \in \mathcal{T}_h$ et un polynôme $r \in \mathbb{P}_p$,

$$\rho_h(u_h, r1_T) = \int_\Omega R_h(u_h) \cdot R_h(r1_T) = \sum_{F \subset \partial T} \int_\Omega R_h(u_h) \cdot r_F(n_T) = \sum_{F \subset \partial T} \int_F n_T \cdot \{R_h(u_h)\},$$

d'où l'on déduit en soustrayant la contribution de ce terme aux flux de LDG et en tenant compte de la relation (87) que le flux pour la méthode SIPG est donné par

$$\phi^u_{T,F}(z_h) = \begin{cases} n_T\cdot(-\{\nabla_h z^u_h\} + \frac{\alpha}{h} n_F [\![z^u_h]\!]), & F \in \mathcal{F}^{\mathrm{i}}_h, \\ n\cdot(-\nabla_h z^u_h + \frac{\alpha}{h} n z^u_h), & F \in \mathcal{F}^{\partial}_h. \end{cases} \tag{107}$$

Ces flux sont, comme on s'y attendait, conservatifs et consistants avec ceux de la solution exacte.

4 Analyse unifiée des méthodes de GD

L'objet de cette section est d'esquisser brièvement les grandes lignes d'une théorie unifiée des méthodes de GD embrassant à la fois l'approximation d'EDPs hyperboliques et elliptiques. La notion clé est celle de *système de Friedrichs*; il s'agit de systèmes d'EDPs d'ordre un équipés d'une propriété de *symétrie* et d'une propriété de *positivité* (ou de *dissipativité*).

Soit un entier $m \geq 1$ (m désigne le nombre d'inconnues dans le système de Friedrichs). Soit Ω un ouvert borné, connexe et lipschitzien de $\mathbb{R}^d$. Un système de Friedrichs se formule à l'aide de la donnée de $(d+1)$ champs à valeurs dans $\mathbb{R}^{m,m}$, $\mathcal{K}$ et $\{\mathcal{A}^k\}_{1\leq k\leq d}$. On suppose que tous ces champs sont dans $[L^\infty(\Omega)]^{m,m}$ et que $\sum_{k=1}^d \partial_k \mathcal{A}^k \in [L^\infty(\Omega)]^{m,m}$ (∂_k désigne la dérivée au sens des distributions dans la k-ième direction spatiale). Les propriétés de symétrie et de positivité (ou de dissipativité) s'expriment sous la forme suivante :

$$\forall k \in \{1,\dots,d\},\ \mathcal{A}^k = (\mathcal{A}^k)^T \ \text{ p.p. dans } \Omega, \tag{A1}$$

$$\exists \mu_0 > 0, \quad \mathcal{K} + \mathcal{K}^T - \sum_{k=1}^d \partial_k \mathcal{A}^k \geq 2\mu_0 \mathcal{I}_m \ \text{ p.p. dans } \Omega, \tag{A2}$$

où $\mathcal{I}_m$ désigne la matrice identité dans $\mathbb{R}^{m,m}$. Par la suite, les inégalités entre matrices (symétriques) dans $\mathbb{R}^{m,m}$ sont à entendre au sens des formes quadratiques associées : pour $\mathcal{G}$ et $\mathcal{H}$ dans $\mathbb{R}^{m,m}$, $\mathcal{G} \geq \mathcal{H}$ signifie que pour tout $\xi \in \mathbb{R}^m$, $(\mathcal{G}\xi,\xi)_{\mathbb{R}^m} \geq (\mathcal{H}\xi,\xi)_{\mathbb{R}^m}$.

On pose $L = [L^2(\Omega)]^m$ que l'on munit du produit scalaire usuel noté $(\cdot,\cdot)_L$. On introduit l'opérateur

$$A : [\mathfrak{C}^1(\Omega)]^m \ni z \longmapsto Az = \sum_{k=1}^d \mathcal{A}^k \partial_k z \in L. \tag{108}$$

Cet opérateur peut être étendu aux fonctions de L telles que la forme linéaire

$$[\mathfrak{D}(\Omega)]^m \ni \phi \longmapsto -\int_\Omega \sum_{k=1}^d z^T \partial_k(\mathcal{A}^k \phi) \in \mathbb{R} \tag{109}$$

est continue sur L; l'objet Az désigne alors la fonction de L qui peut être associée à cet forme linéaire par le biais du théorème de représentation de Riesz–Fréchet. On introduit l'espace du graphe

$$W = \{z \in L;\, Az \in L\}, \tag{110}$$

que l'on munit d'une structure hilbertienne par le biais du produit scalaire $(z,y)_L + (Az,Ay)_L$. Par construction, $A \in \mathcal{L}(W;L)$. On introduit également l'opérateur

$$\tilde{A} : [\mathfrak{C}^1(\Omega)]^m \ni z \longmapsto \tilde{A}z = -\sum_{k=1}^{d} \partial_k(\mathcal{A}^k z) \in L, \tag{111}$$

dont on étend le domaine à W en procédant comme ci-dessus. Enfin, on introduit l'opérateur $D \in \mathcal{L}(W;W')$ en posant

$$\forall (z,y) \in W \times W, \quad \langle Dz, y\rangle_{W',W} = (Az,y)_L - (z,\tilde{A}y)_L. \tag{112}$$

L'opérateur D est auto-adjoint par construction. De plus, D est tel que $[\mathfrak{D}(\Omega)]^m \subset \mathrm{Ker}(D)$; on dit que D est un opérateur de frontière. Par ailleurs, lorsque les champs $\mathcal{A}^k$ sont suffisamment réguliers, on pose

$$\mathcal{D} = \sum_{k=1}^{d} n_k \mathcal{A}^k : \partial\Omega \longrightarrow \mathbb{R}^{m,m}, \tag{113}$$

où n^k est la k-ième composante de la normale extérieure n, et on observe que si z et y sont suffisamment réguliers,

$$\langle Dz, y\rangle_{W',W} = \int_{\partial\Omega} y^T \mathcal{D} z. \tag{114}$$

Ainsi, (112) est une formule d'intégration par parties. Noter que par construction, $\mathcal{D} = \mathcal{D}^T$.

Pour $f \in L$, on cherche $z \in W$ tel que $Kz + Az = f$ où $K : L \ni z \mapsto \mathcal{K}z \in L$ représente le terme d'ordre 0 dans le système de Friedrichs. Afin d'obtenir un problème bien posé, il faut compléter ce système d'EDPS par des conditions aux limites. Pour cela, on suppose qu'il existe un opérateur $M \in \mathcal{L}(W;W')$ tel que

$$\forall z \in W,\ \langle Mz, z\rangle_{W',W} \geq 0, \tag{M1}$$

$$W = \mathrm{Ker}(D-M) + \mathrm{Ker}(D+M). \tag{M2}$$

Observer que M est un opérateur de frontière puisque (M2) implique que $\mathrm{Ker}(D) = \mathrm{Ker}(M)$. Dans la plupart des applications, l'opérateur M peut être représenté sur la frontière par un champ à valeurs dans $\mathbb{R}^{m,m}$, c'est-à-dire qu'il existe $\mathcal{M}$ dans $[L^\infty(\partial\Omega)]^{m,m}$ tel que

$$\langle Mz, y\rangle_{W',W} = \int_{\partial\Omega} y^T \mathcal{M} z. \tag{115}$$

Les conditions aux limites que l'on souhaite imposer s'écrivent sous la forme $z \in \mathrm{Ker}(M-D)$. On définit sur $W \times W$ la forme bilinéaire

$$a(z,y) = (Kz,y)_L + (Az,y)_L + \tfrac{1}{2}\langle (M-D)z, y\rangle_{W',W}, \tag{116}$$

et on considère le problème qui consiste à

$$\text{Chercher } z \in W \text{ tel que } a(z,y) = (f,y)_L,\ \forall y \in W. \tag{117}$$

On a le résultat suivant.

Théorème 21. *Le problème* (117) *est bien posé. De plus, son unique solution est dans* $\mathrm{Ker}(M-D)$ *et elle satisfait* $Kz + Az = f$ *dans* L.

Voici deux exemples de systèmes de Friedrichs.

– L'équation d'advection–réaction (3). On a $m = 1$; on pose

$$\mathcal{K} = \mu \qquad \text{et} \qquad \mathcal{A}^k = \beta^k, \tag{118}$$

où β^k désigne la k-ième composante du champ d'advection β. La propriété de symétrie (A1) est trivialement satisfaite puisque les champs $\mathcal{A}^k$ sont à valeurs scalaires. La propriété de positivité (A2) est une conséquence immédiate de (1). L'espace du graphe est $W = \{v \in L^2(\Omega);\, \beta{\cdot}\nabla v \in L^2(\Omega)\}$. L'opérateur D est tel que

$$\langle Dz, y\rangle_{W',W} = \int_{\partial\Omega} (\beta{\cdot}n) zy, \tag{119}$$

et la condition aux limites $u|_{\partial\Omega^-} = 0$ peut être imposée en considérant l'opérateur M tel que

$$\langle Mz, y\rangle_{W',W} = \int_{\partial\Omega} |\beta{\cdot}n| zy. \tag{120}$$

On vérifie facilement que les hypothèses (M1) et (M2) sont bien satisfaites. Enfin, (114) et (115) sont satisfaites avec $\mathcal{D} = (\beta{\cdot}n)$ et $\mathcal{M} = |\beta{\cdot}n|$.

– L'équation de diffusion–réaction dans sa formulation mixte (54). On a $m = d+1$. Du fait de l'approche par formulation mixte, les champs $\mathcal{K}$ et $\mathcal{A}^k$ ont une structure bloc. On pose

$$\mathcal{K} = \left[\begin{array}{c|c} \mathcal{I}_d & 0 \\ \hline 0 & 1 \end{array}\right] \qquad \text{et} \qquad \mathcal{A}^k = \left[\begin{array}{c|c} 0 & e^k \\ \hline (e^k)^T & 0 \end{array}\right], \tag{121}$$

où $\mathcal{I}_d$ désigne la matrice identité dans $\mathbb{R}^{d,d}$ et e^k le k-ième vecteur de la base canonique de $\mathbb{R}^d$. Les propriétés (A1) et (A2) sont clairement satisfaites. L'espace du graphe est $W = H(\mathrm{div};\Omega) \times H^1(\Omega)$. L'opérateur D est tel que

$$\langle Dz, y\rangle_{W',W} = \langle n{\cdot}y^\sigma, z^u\rangle_{H^{-\frac12},H^{\frac12}} + \langle n{\cdot}z^\sigma, y^u\rangle_{H^{-\frac12},H^{\frac12}}, \tag{122}$$

et la condition aux limites de Dirichlet homogène peut être imposée en considérant l'opérateur M tel que

$$\langle Mz, y\rangle_{W',W} = -\langle n{\cdot}y^\sigma, z^u\rangle_{H^{-\frac12},H^{\frac12}} + \langle n{\cdot}z^\sigma, y^u\rangle_{H^{-\frac12},H^{\frac12}}. \tag{123}$$

On vérifie facilement que les hypothèses (M1) et (M2) sont bien satisfaites. Enfin, (114) et (115) sont satisfaites avec

$$\mathcal{D} = \left[\begin{array}{c|c} 0 & n \\ \hline n^T & 0 \end{array}\right] \qquad \text{et} \qquad \mathcal{M} = \left[\begin{array}{c|c} 0 & -n \\ \hline n^T & 0 \end{array}\right]. \tag{124}$$

D'autres exemples de systèmes de Friedrichs sont l'équation d'advection–diffusion–réaction, les équations de la mécanique des milieux continus et les équations de Maxwell en régime diffusif.

L'approximation GD des systèmes de Friedrichs se construit par le biais de deux familles de champs à valeurs dans $\mathbb{R}^{m,m}$. La première, $\{\mathcal{M}_F\}_{\mathcal{F}_h^\partial}$, est indexée par les faces de frontière ; son rôle est d'imposer les conditions aux limites au sens faible et de contrôler certaines valeurs au bord de la solution discrète. La deuxième famille, $\{\mathcal{S}_F\}_{\mathcal{F}_h^{\rm i}}$, est indexée par les faces intérieures ; son rôle est de pénaliser au sens des moindres carrés le saut à travers les faces intérieures de certaines composantes de la solution discrète. Enfin, il est utile d'étendre le champ $\mathcal{D}$ aux faces intérieures en posant pour chaque $T \in \mathcal{T}_h$, $\mathcal{D}|_{\partial T} = \sum_{k=1}^d n_{T,k}\mathcal{A}^k$ où $n_T = (n_{T,1}, \ldots, n_{T,d})$ désigne la normale unitaire extérieure à T. Le champ $\mathcal{D}$ ainsi défini est bi-valué sur chaque $F \in \mathcal{F}_h^{\rm i}$ avec $\{\mathcal{D}\} = 0$. De plus, les valeurs prises par $\mathcal{D}$ étant des matrices *symétriques* d'ordre m, on peut définir le champ $|\mathcal{D}|$; celui-ci est uni-valué sur chaque $F \in \mathcal{F}_h^{\rm i}$. On suppose que la solution exacte z de (117) est suffisamment régulière pour que :

- $z \in [H^{p+1}(\mathcal{T}_h)]^m$;
- $\mathcal{M}z = \mathcal{D}z$ p.p. sur chaque $F \in \mathcal{F}_h^\partial$;
- $[\![z_h]\!] = 0$ p.p. sur chaque $F \in \mathcal{F}_h^{\rm i}$.

L'approximation GD de z sera cherchée dans $W_h := [\mathbb{G}_h^p]^m$, c'est-à-dire que l'on considère une approximation GD de degré p pour chaque composante de z. On introduit la forme bilinéaire

$$\begin{aligned} a_h(z,y) = &\sum_{T\in\mathcal{T}_h}[(Kz,y)_{L,T} + (Az,y)_{L,T}] + \sum_{F\in\mathcal{F}_h^\partial}\tfrac{1}{2}(\mathcal{M}_F z - \mathcal{D}z, y)_{L,F} \\ &- \sum_{F\in\mathcal{F}_h^{\rm i}} 2(\{\mathcal{D}z\},\{y\})_{L,F} + \sum_{F\in\mathcal{F}_h^{\rm i}}(\mathcal{S}_F[\![z]\!],[\![y]\!])_{L,F}, \end{aligned} \tag{125}$$

la notation $(\cdot,\cdot)_{L,E}$ désignant le produit scalaire usuel de $[L^2(E)]^m$ où E est une maille, une face ou une union de tels objets. Le rôle du troisième terme dans le membre de droite de (125) est d'assurer une propriété de L^2-coercivité discrète pour la forme bilinéaire a_h tout en préservant la consistance du schéma. On vérifie en effet avec les hypothèses ci-dessus pour la solution exacte z que

$$a_h(y_h,y_h) \geq \mu_0\|y_h\|_L^2 + \sum_{F\in\mathcal{F}_h^\partial}(\mathcal{M}_F y_h, y_h)_{L,F} + \sum_{F\in\mathcal{F}_h^{\rm i}}(\mathcal{S}_F[\![y_h]\!],[\![y_h]\!])_{L,F}, \quad \forall y_h \in W_h, \tag{126}$$

$$a_h(z,y_h) = (f,y_h)_L, \quad \forall y_h \in W_h. \tag{127}$$

Le problème discret consiste à

$$\text{Chercher } z_h \in W_h \text{ tel que } a_h(z_h,y_h) = (f,y_h)_L,\ \forall y_h \in W_h. \tag{128}$$

On obtient des estimations d'erreur (quasi-)optimales si les champs $\{\mathcal{M}_F\}_{\mathcal{F}_h^\partial}$ et $\{\mathcal{S}_F\}_{\mathcal{F}_h^{\rm i}}$ satisfont les propriétés suivantes :

$$\mathrm{Ker}(\mathcal{M}-\mathcal{D}) \subset \mathrm{Ker}(\mathcal{M}_F - \mathcal{D}), \tag{DG1A}$$

$$(\mathcal{M}_F v, v)_{L,F} \geq 0, \tag{DG1B}$$

$$|(\mathcal{M}_F v - \mathcal{D}v, w)_{L,F}| \lesssim |v|_{M,F}\|w\|_{L,F}, \tag{DG1C}$$

$$|(\mathcal{M}_F v + \mathcal{D}v, w)_{L,F}| \lesssim \|v\|_{L,F}|w|_{M,F}, \tag{DG1D}$$

$$\mathcal{S}_F = (\mathcal{S}_F)^T \text{ et } |\mathcal{D}| \lesssim \mathcal{S}_F \lesssim \mathcal{I}_m, \tag{DG1E}$$

où $|v|_{M,F}^2 = (\mathcal{M}_F v, v)_{L,F}$. En procédant comme précédemment (condition inf-sup discrète, continuité avec deux normes), on prouve le résultat suivant.

Théorème 22. *On suppose que* $\forall k \in \{1,\dots,k\}$ *et* $\forall T \in \mathcal{T}_h$, $\mathcal{A}^k|_T \in [\mathcal{C}^{0,\frac12}(T)]^{m,m}$. *Dans ces conditions,*

$$\|z - z_h\|_L + \left(\sum_{T\in\mathcal{T}_h} h_T \|A(z-z_h)\|_T^2\right)^{\frac12} \lesssim h^{p+\frac12}\|z\|_{[H^{p+1}(\mathcal{T}_h)]^m}. \tag{129}$$

On obtient ainsi une estimation d'erreur légèrement suboptimale (d'ordre $\frac12$) dans la norme $\|\cdot\|_L$ et une estimation d'erreur optimale dans la norme du graphe brisée.

Pour l'équation d'advection–réaction, on peut choisir

$$\mathcal{M}_F = \beta\cdot n \qquad \text{et} \qquad \mathcal{S}_F = \alpha|\beta\cdot n| \tag{130}$$

pour un paramètre $\alpha > 0$ (dont la valeur peut éventuellement changer de face à face). On vérifie facilement que les propriétés (DG1A)–(DG1E) sont satisfaites.

Pour l'équation de diffusion–réaction, on peut choisir

$$\mathcal{M}_F = \left[\begin{array}{c:c} 0 & -n \\ \hdashline n^T & \eta_1 \end{array}\right] \qquad \text{et} \qquad \mathcal{S}_F = \left[\begin{array}{c:c} \eta_2 n\otimes n & 0 \\ \hdashline 0 & \eta_3 \end{array}\right] \tag{131}$$

pour des paramètres $\eta_1 > 0$, $\eta_2 > 0$ et $\eta_3 > 0$ (dont la valeur peut éventuellement changer de face à face). On vérifie facilement que les propriétés (DG1A)–(DG1E) sont satisfaites.

Les EDPs elliptiques sous forme mixte conduisent à des systèmes de Friedrichs avec une structure bloc particulière. On décompose l'inconnue z en (z^σ, z^u) où z^σ est à valeurs dans $\mathbb{R}^{m_\sigma}$, z^u est à valeurs dans $\mathbb{R}^{m_u}$ et $m = m_\sigma + m_u$. On suppose que

$$\mathcal{K} = \left[\begin{array}{c:c} \mathcal{K}^{\sigma\sigma} & \mathcal{K}^{\sigma u} \\ \hdashline \mathcal{K}^{u\sigma} & \mathcal{K}^{uu} \end{array}\right], \qquad \mathcal{A}^k = \left[\begin{array}{c:c} 0 & \mathcal{B}^k \\ \hdashline (\mathcal{B}^k)^T & \mathcal{C}^k \end{array}\right], \tag{132}$$

le champ $K^{\sigma\sigma}$ étant uniformément défini positif. Observer que la composante z^σ peut être éliminée du système de Friedrichs $Kz + Az = (f^\sigma, f^u)$ puisque

$$z^\sigma = (\mathcal{K}^{\sigma\sigma})^{-1}(f^\sigma - \mathcal{K}^{\sigma u} z^u - Bz^u) \tag{133}$$

où B désigne l'opérateur différentiel tel que $By^u = \sum_{k=1}^d \mathcal{B}^k \partial_k y^u$. Par ailleurs, de par la décomposition bloc (132), il vient, avec des notations évidentes,

$$\mathcal{D} = \left[\begin{array}{c:c} 0 & \mathcal{D}^{\sigma u} \\ \hdashline (\mathcal{D}^{\sigma u})^T & \mathcal{D}^{uu} \end{array}\right]. \tag{134}$$

On suppose que l'on impose une condition aux limites de type Dirichlet par le biais du champ

$$\mathcal{M} = \left[\begin{array}{c:c} 0 & -\mathcal{D}^{\sigma u} \\ \hdashline (\mathcal{D}^{\sigma u})^T & 0 \end{array}\right]. \tag{135}$$

Pour de tels systèmes de Friedrichs, il est possible de construire (comme nous l'avons fait pour le problème de diffusion–réaction dans l'approche à un champ) une approximation GD pour laquelle la composante z_h^σ peut être éliminée en résolvant des problèmes locaux posés sur chaque maille. La clé pour construire une telle approximation avec des estimations d'erreur satisfaisantes est de ne plus pénaliser les sauts de la composante z_h^σ et de pénaliser les sauts et les valeurs à la frontière de la composante z_h^u de manière plus forte. Plus précisément, on suppose que les champs $\{\mathcal{M}_F\}_{\mathcal{F}_h^\partial}$ et $\{\mathcal{S}_F\}_{\mathcal{F}_h^\mathrm{i}}$ satisfont les propriétés suivantes :

$$\mathcal{M}_F = \left[\begin{array}{c:c} 0 & -\mathcal{D}^{\sigma u} \\ \hdashline (\mathcal{D}^{\sigma u})^T & \mathcal{M}_F^{uu} \end{array}\right], \quad \mathcal{S}_F = \left[\begin{array}{c:c} 0 & 0 \\ \hdashline 0 & \mathcal{S}_F^{uu} \end{array}\right], \tag{DG2A}$$

$$\mathcal{M}_F^{uu} = (\mathcal{M}_F^{uu})^T \quad \text{et} \quad h_F^{-1}((\mathcal{D}^{\sigma u})^T\mathcal{D}^{\sigma u})^{\frac{1}{2}} + h_F|\mathcal{D}^{uu}| \lesssim \mathcal{M}_F^{uu} \lesssim h_F^{-1}\mathcal{I}_{m_u}, \tag{DG2B}$$

$$\mathcal{S}_F^{uu} = (\mathcal{S}_F^{uu})^T \text{ et } h_F^{-1}((\mathcal{D}^{\sigma u})^T\mathcal{D}^{\sigma u})^{\frac{1}{2}} + h_F|\mathcal{D}^{uu}| \lesssim \mathcal{S}_F^{uu} \lesssim h_F^{-1}\mathcal{I}_{m_u}. \tag{DG2C}$$

On a le résultat suivant.

Théorème 23. *On suppose que* $\forall k \in \{1,\dots,k\}$ *et* $\forall T \in \mathcal{T}_h$, $\mathcal{B}^k|_T \in [\mathcal{C}^{0,1}(T)]^{m_\sigma,m_u}$. *Dans ces conditions,*

$$\|z - z_h\|_L + \left(\sum_{T\in\mathcal{T}_h} \|B(z^u - z_h^u)\|_T^2\right)^{\frac{1}{2}} \lesssim h^p \|z\|_{[H^p(\mathcal{T}_h)]^{m_\sigma}\times[H^{p+1}(\mathcal{T}_h)]^{m_u}}. \tag{136}$$

On obtient une estimation d'erreur suboptimale (d'ordre 1) pour la norme $\|\cdot\|_L$ et une estimation d'erreur optimale pour la norme du graphe brisée associée à la composante z^u. Une estimation d'erreur optimale en norme L^2 pour la composante z^u peut être prouvée sous les hypothèses de régularité usuelles grâce à des techniques de dualité.

Pour l'équation de diffusion–réaction, on peut choisir

$$\mathcal{M}_F = \left[\begin{array}{c:c} 0 & -n \\ \hdashline n^T & \frac{\eta_1}{h_F} \end{array}\right] \qquad \text{et} \qquad \mathcal{S}_F = \left[\begin{array}{c:c} 0 & 0 \\ \hdashline 0 & \frac{\eta_3}{h_F} \end{array}\right]. \tag{137}$$

On vérifie facilement que les propriétés (DG2A)–(DG2C) sont satisfaites.

5 Notes bibliographiques

Les méthodes de GD ont été introduites au début des années 70 pour la simulation de problèmes de neutronique [23]. L'analyse numérique des méthodes de GD pour les équations hyperboliques remonte aux travaux de Lesaint et Raviart [19, 20] dans le milieu des années 70 ; l'estimation d'erreur en norme L^2 d'ordre $p+\frac{1}{2}$ a été obtenue dans les années 80 par Johnson et collaborateurs [18]. Les méthodes de GD ont connu un regain d'intérêt considérable pour la simulation d'équations et de systèmes d'équations hyperboliques depuis une dizaine d'années. En particulier, les liens entre méthodes de volumes finis et méthodes de GD ont permis de transposer à ces dernières un certain nombre de techniques bien rodées dans le cadre des méthodes de volumes finis, comme la notion de flux numériques, de solveurs de Riemann approchés et de limiteurs de pente ; voir par exemple [8] pour une revue.

Le développement de méthodes de GD pour approcher la solution d'EDPs elliptiques remonte aux travaux de Nitsche [21] pour imposer des conditions aux limites de Dirichlet au sens faible et aux travaux de Babuška [3], Babuška et Zlámal [4], Douglas et Dupont [10], Baker [5] et Arnold [1] pour l'utilisation de techniques de pénalisation. En particulier, les travaux de Baker puis d'Arnold à la fin des années 70 sont à l'origine de la méthode connue sous le nom de SIPG. La variante (indiquée ci-dessus sous le nom de méthode BRMPS) où les termes de pénalisation sont formulés à l'aide de l'opérateur de relèvement a été introduite par Bassi, Rebay et collaborateurs [7]. Par ailleurs, l'approximation par des méthodes de GD de la forme mixte d'équations elliptiques a été proposée par Bassi et Rebay [6] puis étendue par Cockburn et Shu [9] donnant ainsi naissance à la fin des années 90 à la méthode connue sous le nom de LDG. Enfin, pour les variantes non symétriques des méthodes de GD pour les problèmes elliptiques, voir en particulier [22] et [24]. Une analyse unifiée des méthodes de GD pour le problème de Poisson a été proposée en 2001 par Arnold, Brezzi, Cockburn et Marini [2].

L'analyse unifiée des méthodes de GD embrassant les EDP hyperboliques et elliptiques est due aux travaux récents de Ern et Guermond sur l'approximation des systèmes de Friedrichs par des méthodes de GD [12, 13, 14]. La notion de systèmes d'EDP d'ordre un symétriques et dissipatifs a été introduite par Friedrichs en 1958 [16]. Des développements récents sur l'analyse mathématique de tels systèmes peuvent être trouvés dans [15] et [17].

Enfin, l'analyse de convergence des méthodes d'éléments finis basée sur une condition inf-sup discrète, une propriété de continuité et l'utilisation de deux normes est détaillée par exemple dans [11] où elle est appliquée à de nombreux exemples.

Références

[1] D.N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19 :742–760, 1982.

[2] D.N. Arnold, F. Brezzi, B. Cockburn, and L.D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5) :1749–1779, 2001/02.

[3] I. Babuška. The finite element method with penalty. *Math. Comp.*, 27 :221–228, 1973.

[4] I. Babuška and M. Zlámal. Nonconforming elements in the finite element method with penalty. *SIAM J. Numer. Anal.*, 10(5) :863–875, 1973.

[5] G.A. Baker. Finite element methods for elliptic equations using nonconforming elements. *Math. Comp.*, 31(137) :45–59, 1977.

[6] F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. *J. Comput. Phys.*, 131(2) :267–279, 1997.

[7] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini. A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In R. Decuypere and G. Dibelius, editors, *Proceedings of 2nd European Conference on Turbomachinery, Fluid Dynamics and Thermodynamics*, pages 99–108, Antwerpen, Belgium, 1997. Technologisch Instituut.

[8] B. Cockburn, G.E. Karniadakis, and C.W. Shu. *Discontinuous Galerkin Methods - Theory, Computation and Applications*, volume 11 of *Lecture Notes in Computer Science and Engineering*. Springer, 2000.

[9] B. Cockburn and C.W. Shu. The local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM J. Numer. Anal.*, 35 :2440–2463, 1998.

[10] J. Douglas Jr. and T. Dupont. *Interior Penalty Procedures for Elliptic and Parabolic Galerkin Methods*, volume 58 of *Lecture Notes in Physics*. Springer-Verlag, Berlin, 1976.

[11] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, NY, 2004.

[12] A. Ern and J.-L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. I. General theory. *SIAM J. Numer. Anal.*, 44(2) :753–778, 2006.

[13] A. Ern and J.-L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. II. Second-order PDEs. *SIAM J. Numer. Anal.*, 2006.

[14] A. Ern and J.-L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. III. Multifield theory with partial coercivity. *SIAM J. Numer. Anal.*, 2007. To appear.

[15] A. Ern, J.-L. Guermond, and G. Caplain. An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrichs' systems. *Comm. Partial. Differ. Equ.*, 2007. To appear.

[16] K.O. Friedrichs. Symmetric positive linear differential equations. *Comm. Pure Appl. Math.*, 11 :333–418, 1958.

[17] M. Jensen. *Discontinuous Galerkin Methods for Friedrichs Systems with Irregular Solutions*. PhD thesis, University of Oxford, 2004.

[18] C. Johnson, U. Nävert, and J. Pitkäranta. Finite element methods for linear hyperbolic equations. *Comput. Methods Appl. Mech. Engrg.*, 45 :285–312, 1984.

[19] P. Lesaint. *Sur la résolution des systèmes hyperboliques du premier ordre par des méthodes d'éléments finis*. PhD thesis, University of Paris VI, 1975.

[20] P. Lesaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 89–123. Publication No. 33. Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974.

[21] J. Nitsche. Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg*, 36 :9–15, 1971.

[22] J.T. Oden, I. Babuška, and C.E. Baumann. A discontinuous *hp* finite element method for diffusion problems. *J. Comput. Phys.*, 146(2) :491–519, 1998.

[23] W.H. Reed and T.R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[24] B. Rivière, M.F. Wheeler, and V. Girault. Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. I. *Comput. Geosci.*, 8 :337–360, 1999.