Finite Element approximation of nonlinear conservation problems

J.-L. Guermond

LIMSI (CNRS-UPR 3251), BP 133, F-91403, Orsay, e-mail:`guermond@limsi.fr`

Abstract. The goal of this paper is to present a stabilized Galerkin technique for approximating non-coercive PDE's. This technique is based on a two-level hierarchical decomposition of the approximation space. This space is broken up into resolved scales and subgrid scales. We show that in general the Galerkin formulation provides an a priori control on the resolved scales of the approximate solution, whereas it cannot control the subgrid scales. The missing stability is obtained by slightly modifying the Galerkin formulation by introducing an artificial diffusion on the subgrid scales. Numerical tests show that the method applies also to nonlinear problems.

1 Model problems

In this section we recall an abstract existence result and we show that the Galerkin formulation is not optimal for approximating PDE's dominated by first order differential operators.

1.1 An abstract existence and stability result

Let $V \subset L$ be two real Hilbert spaces with dense and continuous embedding. For any Hilbert space H we denote by $(\cdot, \cdot)_H$ and $\|\cdot\|_H$ the scalar product and the norm in H respectively. For any Banach space, we denote by $B' = \mathcal{L}(B; \mathbb{R})$ the dual of B. Hereafter we make the usual identifications $V \subset L \equiv L' \subset V'$. Let $a \in \mathcal{L}(V \times L; \mathbb{R})$ and consider the following problem.

$$\begin{cases} \text{For all } f \in L, \text{ find } u \in V \text{ s.t.} \\ a(u, v) = (f, v)_L, \qquad \forall v \in L. \end{cases} \quad (1.1)$$

Sufficient and necessary conditions for this problem to be well-posed are stated in the following theorem due to Nečas [Neč62]:

Theorem 1.1 (Nečas). *Problem 1.1 is well-posed if and only if*

$$\exists \alpha > 0, \qquad \inf_{u \in V} \sup_{v \in L} \frac{a(u, v)}{\|u\|_V \|v\|_L} \geq \alpha, \quad (1.2)$$

$$\forall v \in L, \quad (v \neq 0) \Rightarrow (\sup_{u \in V} \frac{a(u, v)}{\|u\|_V} \neq 0). \quad (1.3)$$

To interpret this theorem, let us define the operator $A : D(A) = V \subset L \longrightarrow L$ such that $(Au, v)_L = a(u, v)$ for all $(u, v) \in V \times L$. Condition (1.2) is equivalent to assuming that A is injective and its range is closed, whereas (1.3) states that A^t is injective. As a result, these two conditions are equivalent to assuming that A is bijective [Bre91].

Now let us look at the approximation of (1.1). Let $V_h \subset V$ and $L_h \subset L$ be two finite-dimensional vector spaces and consider the following discrete problem.

$$\begin{cases} \text{Find } u_h \in V_h \text{ s.t.} \\ a(u_h, v_h) = (f, v_h)_L, \qquad \forall v_h \in L_h. \end{cases} \quad (1.4)$$

Proposition 1.1. *Assume that* $\dim V_h = \dim L_h$ *and there is* $\alpha_h > 0$ *such that for all* $w_h \in V_h$

$$\sup_{v_h \in L_h} \frac{a(w_h, v_h)}{\|v_h\|_L} \geq \alpha_h \|w_h\|_V. \quad (1.5)$$

Then, problem (1.4) has a unique solution and $\|u_h\|_V \leq \frac{1}{\alpha_h}\|f\|_L$.

Lemma 1.1 (Céa). *Under the hypotheses of theorem 1.1 and proposition 1.1 we have*

$$\|u - u_h\|_V \leq (1 + \frac{\|a\|}{\alpha_h}) \inf_{w_h \in W_h} \|u - w_h\|_V. \quad (1.6)$$

1.2 Example 1 : advection/reaction

Let us consider an advection/reaction problem. Let β be a smooth vector field in $\mathbb{R}^d$, say $\beta \in L^\infty(\Omega)^d$ and $\nabla \cdot \beta \in L^\infty(\Omega)$, and set

$$\Gamma^- = \{x \in \Gamma \mid \beta(x) \cdot n(x) < 0\},$$
$$\Gamma^+ = \{x \in \Gamma \mid \beta(x) \cdot n(x) > 0\}.$$

Γ^- is the inflow boundary and Γ^+ is the outflow boundary. It may happen that these two subsets of Γ are empty if β is such that $\beta \cdot n(x) = 0$ for all $x \in \Gamma$. Let μ be a function in $L^\infty(\Omega)$. We introduce the following differential operator

$$A(u) = \mu u + \beta \cdot \nabla u.$$

To give a precise meaning to A, we introduce its domain

$$V = D(A) = \{w \in L^2(\Omega); \ \beta \cdot \nabla w \in L^2(\Omega)\} \subset L^2(\Omega).$$

When equipped with the norm $\|w\|_V = (\|w\|_{0,\Omega}^2 + \|\beta \cdot \nabla w\|_{0,\Omega}^2)^{1/2}$, it is clear that V is a Hilbert space and $A \in \mathcal{L}(V; L)$. In general A is not an isomorphism if we do not assume any other hypotheses on μ and β. Hereafter we assume that there is $\mu_0 > 0$ so that

$$\mu(x) - \frac{1}{2}\nabla \cdot \beta(x) \geq \mu_0 > 0 \quad \text{a.e. } x \text{ in } \Omega. \qquad (1.7)$$

We define $V_0 = \{w \in V; \ w_{|\Gamma^-} = 0\}$. We introduce the bilinear form $a \in \mathcal{L}(V_0 \times L^2(\Omega); \mathbb{R})$ associated with the restriction of A to V_0 :

$$\begin{aligned}\forall u \in V_0, \ \forall v \in L^2(\Omega), \\ a(u,v) = (\mu u + \beta \cdot \nabla u, v)_{0,\Omega}.\end{aligned} \qquad (1.8)$$

Lemma 1.2. *The bilinear form defined in (1.8) satisfies the two conditions of the Nečas theorem.*

The consequence of this lemma is that for all $f \in L^2(\Omega)$, the following problem

$$\begin{cases} \text{Find } u \text{ in } V_0 \text{ s.t.} \\ a(u,v) = (f,v)_{0,\Omega}, \qquad \forall v \in L^2(\Omega), \end{cases} \qquad (1.9)$$

has a unique solution. Equivalently, it means that $A : V_0 \longrightarrow L^2(\Omega)$ is an isomorphism.

Remark 1.1. If $\mu = 0$ and $\nabla \cdot \beta = 0$, the hypothesis (1.7) is not satisfied. Nevertheless, the conclusions of lemma 1.2 still hold if β is a filling field: i.e., if for almost every x in Ω, there is a characteristics of β that starts from x and reaches Γ^- in finite time. The reader is referred to Azerad and Pousin [AP96] for other details on this problem.

1.3 Example 2 : The Darcy equation

let Ω be a porous medium characterized by the permeability tensor $K(x)$. This tensor is assumed to be symmetric positive definite and its smallest and largest eigen values are assumed to be bounded from below and from above uniformly in Ω. Let $\Gamma = \Gamma_1 \cup \Gamma_2$ be a partition of Γ. We consider the following problem:

$$\begin{cases} K^{-1} \cdot u + \nabla p = f \\ \nabla \cdot u = g \\ u \cdot n_{|\Gamma_1} = 0, \quad p_{|\Gamma_2} = 0. \end{cases} \qquad (1.10)$$

This problem is known as the Darcy problem. In nonlinear form, it plays an important role in underground storage problems, hydro-geology, and in the petroleum industry. It is very often coupled to a transport equation for the concentration of a chemical specie or a phase fraction.

To formulate (1.10) in weak form, we introduce some definitions.

$$X = \{v \in L^2(\Omega)^d; \ \nabla \cdot v \in L^2(\Omega), \ v \cdot n_{|\Gamma_1} = 0\},$$
$$\|v\|_X = (\|v\|_{0,\Omega}^2 + \|\nabla \cdot v\|_{0,\Omega}^2)^{1/2},$$
$$Y = \{q \in L^2(\Omega); \ \nabla q \in L^2(\Omega), \ q_{|\Gamma_2} = 0\},$$
$$\|q\|_Y = \|q\|_{1,\Omega}.$$

X and Y are Hilbert spaces. We set $V = X \times Y$ and $L = L^2(\Omega)^d \times L^2(\Omega)$ that we equip with the norms $\|(v,q)\|_V = (\|v\|_X^2 + \|q\|_Y^2)^{1/2}$ and $\|(v,q)\|_L = (\|v\|_{0,\Omega}^2 + \|q\|_{0,\Omega}^2)^{1/2}$ respectively. We now define the operator

$$A : V \longrightarrow L$$
$$(v,q) \longmapsto (K^{-1}v + \nabla q, \nabla \cdot v).$$

A is clearly continuous. Finally, we introduce the bilinear form $a \in \mathcal{L}(V \times L; \mathbb{R})$ such that $a((u,p),(v,q)) = (A(u,p),(v,q))_L$.

Lemma 1.3. *The bilinear form a satisfies the two conditions of the Nečas theorem.*

The direct consequence of this lemma is that for all $f \in L^2(\Omega)^d$ and $q \in L^2(\Omega)$, the following problem

$$\begin{cases} \text{Find } (u,p) \in V \text{ s.t. } \forall (v,q) \in L \\ a((u,p),(v,q)) = ((f,g),(v,q))_L, \end{cases} \qquad (1.11)$$

has a unique solution.

1.4 A 1D model problem

Let us simplify the advection problem (1.9). Let $\Omega =]0,1[$ and set $\beta = 1$, $\mu = 0$. We define the Hilbert space $X = \{v \in H^1(\Omega); \ v(0) = 0\}$, and we set

$$a(u,v) = \int_0^1 u'(x)v(x).$$

It is clear that $a \in \mathcal{L}(X \times L^2(\Omega); \mathbb{R})$ and a satisfies the hypotheses of theorem 1.1. In this section we shall consider the following problem. For $f \in L^2(\Omega)$

$$\begin{cases} \text{Find } u \text{ in } X \text{ s.t.} \\ a(u,v) = \int_0^1 fv, \qquad \forall v \in L^2(\Omega). \end{cases} \qquad (1.12)$$

This problem has a unique solution in X. We shall now build a Galerkin approximation of u by means of $\mathbb{P}_1$ finite elements, and we shall see that this approach is not optimal.

Let us define a mesh on $\overline{\Omega} = [0,1]$. For $N \in \mathbb{N}^\star$ set $h = 1/N$ and $x_i = ih$ for $i \in \{0, 1, \ldots, N\}$. We define

$$\begin{aligned}X_h = \{v_h \in \mathcal{C}^0(\overline{\Omega}); \ v_{h|[x_i, x_{i+1}]} \in \mathbb{P}_1, \\ 0 \leq i \leq N-1; \ v_h(0) = 0\}.\end{aligned} \qquad (1.13)$$

It is clear that $X_h \subset X$. The discrete Galerkin formulation of (1.12) is

$$\begin{cases} \text{Find } u_h \text{ in } X_h \text{ s.t.} \\ a(u_h, v_h) = \int_0^1 f v_h, \qquad \forall v_h \in X_h. \end{cases} \quad (1.14)$$

We can apply proposition 1.1 with $V_h = L_h = X_h$ The discrete problem is well posed iff there is $\alpha_h > 0$ such that (1.5) holds. Furthermore, the error estimate is optimal only if α_h is uniformly bounded from below as $h \to 0$. Unfortunately we can prove the following negative theorem.

Theorem 1.2. *There are two constants $c_1 > 0$ and $c_2 > 0$, independent of h, s.t.*

$$c_1 h \leq \inf_{u_h \in X_h} \sup_{v_h \in X_h} \frac{a(u_h, v_h)}{\|u_h\|_{1,\Omega} \|v_h\|_{0,\Omega}} \leq c_2 h.$$

Proof. let us assume that N is even; the other case can be treated similarly. Let $(\phi_i)_{1 \leq i \leq N}$ be the base of X_h such that $\phi_i(x_j) = \delta_{ij}$. To prove the bound from above let us consider the following oscillating function u_h:

$$u_h = \sum_1^N U_i \phi_i \quad \text{with} \quad \begin{cases} U_{2i} = 2ih, & \text{if } 1 \leq i \leq \frac{N}{2}, \\ U_{2i+1} = 1, & \text{if } 0 \leq i \leq \frac{N}{2} - 1. \end{cases}$$

Let f_h be the L^2 projection of u_h' onto X_h. After some calculus we can show

$$\sup_{v_h \in X_h} \frac{a(u_h, v_h)}{\|u_h\|_{1,\Omega} \|v_h\|_{0,\Omega}} \leq \frac{1}{|u_h|_{1,\Omega}} \sup_{v_h \in X_h} \frac{\int_0^1 f_h v_h}{\|v_h\|_{0,\Omega}}$$
$$= \frac{\|f_h\|_{0,\Omega}}{|u_h|_{1,\Omega}} \leq 4\sqrt{3} h.$$

In other words, the quantity $|u_h|_{1,\Omega}$ diverges when $h \to 0$ whereas the L^2 projection of u_h' onto X_h is bounded. The bound from below is evident. ∎

Remark 1.2. The consequence of this negative theorem is that the Galerkin technique is not optimal for approximating first order PDE's.

2 Stabilization by means of a subgrid viscosity method

In this section, we propose a new technique that is optimal for approximating first order PDE's. This technique is based on a hierarchical two-level decomposition of the approximation space. Hereafter we assume that a is positive; i.e.

$$\forall v \in V, \qquad a(v, v) \geq 0.$$

2.1 Introduction

To build an approximate solution to problem (1.1), we introduce a sequence of finite dimensional spaces

$$(X_H)_{(H>0)} \subset V,$$

and we assume that there is a dense subspace $W \subset V$ together with a linear interpolation operator $I_H \in \mathcal{L}(W; X_H)$ and two constants $k > 0$, $c > 0$ such that for all $H > 0$ and all v in W

$$\|v - I_H v\|_L + H\|v - I_H v\|_V \leq cH^{k+1}\|v\|_W. \quad (2.1)$$

Theorem 1.2 clearly states that (1.1) cannot be approximated by means of the Galerkin technique when A is a first order differential operator. A simple cure to this problem consists of enlarging the space of the test functions. Indeed, it is clear that

$$\inf_{u_H \in X_H} \sup_{v \in L} \frac{a(u_H, v)}{\|u_H\|_V \|v\|_L} \geq \alpha.$$

Hence, it is likely that, still approximating u in X_H, there exists a discrete space wedged between X_H and L such that the inf-sup inequality (1.5) is satisfied uniformly. For the time being, let us denote by X_h this space, and let us assume that there is a finite-dimensional space $X_h^H \subset V \subset L$ with $X_h^H \cap X_H = \emptyset$, such that $X_h = X_H \oplus X_h^H$ and the bilinear form a satisfies uniformly the discrete inf-sup inequality, i.e., there is $c_a > 0$, s.t. for all $H > 0$, $h > 0$

$$\inf_{v_H \in X_H} \sup_{\phi_h \in X_h} \frac{a(v_H, \phi_h)}{\|v_H\|_V \|\phi_h\|_L} \geq c_a. \quad (2.2)$$

We are now in measure of building a Petrov–Galerkin approximation:

$$\begin{cases} \text{Find } u_H \in X_H, \text{ s.t.} \\ a(u_H, v_h) = (f, v_h)_L, \qquad v_h \in X_h. \end{cases}$$

Clearly, if this problem has a solution, (2.2) states that this solution is stable uniformly with respect to H and h. Unfortunately, the dimension of X_h is larger than that of X_H; as a result, proposition 1.1 does not hold. To avoid this dimension problem, we could try to approximate u in X_h and test the equation with X_h. By doing so we would be led back to the Galerkin formulation, which we know is not optimal in general. Let us summarize:

1. The hypothesis (2.2) allows for a control on u_H.
2. To have as many unknown as equations, we want to work with one discrete space only, but the Galerkin formulation cannot control correctly u_h. That is to say, we have no a priori control on the quantity $u_h - u_H$.
3. One simple way to control $u_h - u_H$ is to add to our problem a coercive bilinear form acting only on $u_h - u_H$ and small enough such that it does not spoil the consistency.

Let us now state precisely the hypotheses that we need to carry out our program.

1. We assume the decomposition $X_h = X_H \oplus X_h^H$ to be direct. We define $P_H : X_h \longrightarrow X_H$ as being the projection of X_h onto X_H that is parallel to X_h^H. We assume that P_H is stable in the norm of L uniformly with respect to H and h. For all v_h in X_h we denote

$$v_H = P_H v_h \qquad \text{and} \qquad v_h^H = (1 - P_H) v_h. \quad (2.3)$$

2. X_h being finite dimensional, we assume that there is $c_i > 0$, independent of h and H, s.t.

$$\forall v_h \in X_h, \qquad \|v_h\|_V \leq c_i H^{-1} \|v_h\|_L. \qquad (2.4)$$

3. We introduce a norm $\|\cdot\|_b$ s.t.

$$\exists c_{e1} > 0, \ \exists c_{e2} > 0, \ \forall v_h^H \in X_h^H,$$
$$c_{e1}\|v_h^H\|_V \leq \|v_h^H\|_b \leq c_{e2} H^{-1} \|v_h^H\|_L. \qquad (2.5)$$

4. We define a bilinear form $b_h \in \mathcal{L}(X_h^H \times X_h^H; \mathbb{R})$ such that for all (v_h^H, w_h^H) in $X_h^H \times X_h^H$

$$c_b H \|v_h^H\|_b^2 \leq b_h(v_h^H, v_h^H) \quad \text{and}$$
$$b_h(v_h^H, w_h^H) \leq c_{b2} H \|v_h^H\|_b \|w_h^H\|_b. \qquad (2.6)$$

Remark 2.1. We shall hereafter refer to X_H and X_h^H as the resolved scales space and the subgrid scales space respectively. The operator P_H can be thought of as a filter which when acting on a function of X_h gets rid of its fluctuating subgrid scales.

Remark 2.2. The property (2.4) is an inverse inequality. Note that for finite elements, if A is a first order differential operator and h, H denote the mesh size on which X_h and X_H are built respectively, then (2.4) holds true if H and h are of the same order; i.e., $c_1 h \leq H \leq c_2 h$. In practice we shall always use $H = 2h$.

Remark 2.3. Let us give some examples. Assume that $(\cdot, \cdot)_V$ is the scalar product of V. The simplest choice for b_h consists in $b_h(v_h^H, w_h^H) = H(v_h^H, w_h^H)_V$. For the advection problem of section 1.2, one can choose

$$b_h(v_h^H, w_h^H) = H(\mu v_h^H + \beta \cdot \nabla v_h^H, \mu w_h^H + \beta \cdot \nabla w_h^H)_{0,\Omega}.$$

Within this framework we have $\|\cdot\|_b = \|\cdot\|_V$. There is a second possibility if one can exhibit a subspace $X \subset V$ with dense and continuous embedding such that the following inverse inequality holds: $\|v_h\|_X \leq c_{e2} H^{-1}\|v_h\|_L$ for all v_h^H in X_h^H. In practice, this hypothesis means that V and X are domains of differential operators of the same order. Assume that $X_h \subset X$, and denote by $(\cdot, \cdot)_X$ the scalar product in X. One can set $b_h(v_h^H, w_h^H) = H(v_h^H, w_h^H)_X$. For the advection problem of section 1.2, we have $X = H_0^1(\Omega) \subset V$, and assuming $X_h^H \subset H_0^1(\Omega)$ we can set

$$b_h(v_h^H, w_h^H) = H(v_h^H, w_h^H)_{0,\Omega} + H(\nabla v_h^H, \nabla w_h^H)_{0,\Omega}.$$

In this case we have $\|\cdot\|_b = \|\cdot\|_{1,\Omega}$. In practice, this bilinear form is simple to program and problem independent.

Remark 2.4. To some extent, the idea of scale separation and subgrid viscosity is rooted in the spectral viscosity theory developed by Tadmor [Tad89] for approximating nonlinear conservation laws by means of spectral methods.

2.2 The discrete problem

The discrete problem we consider now reads:

$$\begin{cases} \text{Find } u_h \in X_h \text{ s.t. } \forall v_h \in X_h \\ a(u_h, v_h) + b_h(u_h^H, v_h^H) = (f, v_h)_L. \end{cases} \qquad (2.7)$$

Remark 2.5. Note that the only difference between the Galerkin formulation and (2.7) consists of the presence of the bilinear form b_h; i.e., for the Galerkin formulation $b_h = 0$.

Let us define $a_s(u, v) = \frac{1}{2}(a(u, v) + a(v, u))$. It is clear that $a_s \in \mathcal{L}(V \times V; \mathbb{R})$ and a_s is symmetric positive. The major result of this section is the following.

Theorem 2.1. *Under the hypotheses (2.1) to (2.6), problem (2.7) has a unique solution u_h, and if, u, the solution to (1.1) is in W we have the following error estimates.*

$$\begin{cases} a_s(u - u_h, u - u_h)^{1/2} \leq c H^{k+1/2} \|u\|_W, \\ \|u - u_h\|_V + \|u_h^H\|_b \leq c H^k \|u\|_W. \end{cases} \qquad (2.8)$$

Proof. See Guermond [Gue99b, Gue99a]. □

Remark 2.6. The estimate (2.8) is optimal in V. If a_s is L-coercive, (2.8) is not optimal in L; a factor $H^{1/2}$ is missing. Optimality can recovered for finite elements if the mesh satisfies special geometric properties (see [Zho97] for details).

Remark 2.7. The estimate (2.8) is identical to that obtained with the Galerkin Least Square method [JNP84].

2.3 Refinement of the hypotheses

It happens frequently that the operator A can be decomposed into $A = A_0 + A_1$ where A_0 is a zeroth order operator and A_1 is a first order differential operator. For instance, for the advection operator considered in section 1.2, we have $A_0 u = \mu u$ and $A_1 u = \beta \cdot \nabla u$.

Let us consider the decomposition $a = a_0 + a_1$ where $a_0(u, v) = (A_0 u, v)_L$ and $a_1(u, v) = (A_1 u, v)_L$. We now make the following hypotheses:

1. There is a semi-norm in V, which we denote by $|\cdot|_V$, such that the decomposition $a = a_0 + a_1$ satisfies:

$$\forall (u, v) \in V \times L \begin{cases} \|u\|_V \leq c(a_s(u, u)^{1/2} + |u|_V), \\ a_0(u, v) \leq c_0 a_s(u, u)^{1/2} \|v\|_L, \\ a_1(u, v) \leq c_1 |u|_V \|v\|_L. \end{cases} \qquad (2.9)$$

2. We weaken hypothesis (2.2) by replacing it by: there are two constants $c_{a1} > 0$, $c_\delta \geq 0$, independent of (H, h), s.t. for all u_h in X_h

$$\sup_{v_h \in X_h} \frac{a_1(u_H, v_h)}{\|v_h\|_L} \geq c_{a1} |u_H|_V - c_\delta a_s(u_h, u_h)^{1/2}. \qquad (2.10)$$

3. We weaken the definition of b_h. We assume that there is a semi-norm $|\cdot|_b$ such that b_h satisfies the following properties: for all $(v_h^H, w_h^H) \in X_h^H \times X_h^H$,

$$\begin{cases} c_{e1}|v_h^H|_V \leq |v_h^H|_b \leq c_{e2}H^{-1}\|v_h^H\|_L, \\ b_h(v_h^H, v_h^H) \geq c_{b1}H|v_h^H|_b^2, \\ b_h(v_h^H, w_h^H) \leq c_{b2}H|v_h^H|_b|w_h^H|_b. \end{cases} \quad (2.11)$$

Remark 2.8. The reason for weakening (2.2) is that (2.10) is usually simpler to prove.

Remark 2.9. For the advection equation $\mu u + \beta \cdot \nabla u = f$, assuming $\mu - \frac{1}{2}\nabla \cdot \beta \geq \mu_0 > 0$, the bilinear form a is $L^2(\Omega)$-coercive. Hence, one can use the following definition

$$\begin{aligned} \forall (v_h^H, w_h^H) \in X_h^H \times X_h^H \\ b_h(v_h^H, w_h^H) = H(\nabla v_h^H, \nabla w_h^H)_{0,\Omega}. \end{aligned} \quad (2.12)$$

Proposition 2.1. *Under the hypotheses (2.9), (2.1), (2.3), (2.4), (2.10) and (2.11), if the solution to (1.1) is in W, the solution to (2.7) satisfies the estimates (2.8).*

Remark 2.10. The theory developed above generalizes easily to non-uniform regular meshes provided the definition of the bilinear form b_h is localized, see [Gue99a] and [Gue01b].

2.4 A singular perturbation problem

The technique developed above is tailored for problems where A is a first order differential operator. In practice we frequently have to deal with operators of the form $B = A + \epsilon D$, where A is a positive first order differential operator and D is second order and coercive. Given the positiveness of A, the operator B is coercive with ϵ as the coercivity constant. If ϵ is of order 1, the problem $Bu = f$ is elliptic and can easily be approximated by means of the Galerkin technique. On the other hand, if ϵ is small the coercivity is not strong enough to guarantee the Galerkin technique to work properly, for in first approximation $B \approx A$. We shall show in the following that the subgrid viscosity technique developed above generalizes to this situation and yield optimal convergence estimates.

Let us retain the same hypotheses on a, V, and L as before. Moreover, we introduce a new Hilbert space X, and we assume that $X \subset V$ with dense and continuous embedding. We define $d \in \mathcal{L}(X \times X; \mathbb{R})$ and we assume that the bilinear form $a + d$ is X-coercive, i.e., $\|v\|_X^2 \leq a(v,v) + d(v,v)$. For $0 \leq \epsilon \leq 1$, we consider the following problem. For $f \in L$,

$$\begin{cases} \text{Find } u \in X \text{ s.t.} \\ a(u,v) + \epsilon d(u,v) = (f,v), \quad \forall v \in X. \end{cases} \quad (2.13)$$

Remark 2.11. For an advection/diffusion/reaction problem we have $a(u,v) = (\mu u + \beta \cdot \nabla u, v)_{0,\Omega}$ and $d(u,v) = (\nabla u, \nabla v)_\Omega$ with $X = H_0^1(\Omega)$, $V = \{v \in L^2(\Omega); \beta \cdot \nabla v \in L^2(\Omega); w_{|\Gamma^-} = 0\}$, and $L = L^2(\Omega)$.

Let us now approximate the solution to problem (2.13). Let $X_h \subset X_H \subset X$ satisfying hypotheses (2.1), (2.3), (2.4), (2.9), (2.10), and (2.11). Assume furthermore that there is $c > 0$ independent of (H, h) such that

$$\forall v_h \in X_h, \qquad \|v_h\|_X \leq cH^{-1}\|v_h\|_L. \quad (2.14)$$

This hypothesis means that X and V are domains of differential operators of the same order. The discrete problem with consider now reads:

$$\begin{cases} \text{Find } u_h \in X_h \text{ s.t. } \forall v_h \in X_h \\ a(u_h, v_h) + \epsilon d(u_h, v_h) + b_h(u_h^H, v_h^H) = (f, v_h). \end{cases} \quad (2.15)$$

Theorem 2.2. *Under the hypotheses (2.1), (2.3), (2.4), (2.9), (2.10), (2.11) and (2.14), and provided that $u \in W$, the solution to (2.15) satisfies*

$$\begin{cases} a_s(u - u_h, u - u_h)^{1/2} + \epsilon^{1/2}\|u - u_h\|_X \\ \qquad \leq c(H^{k+1/2} + H^k\epsilon^{1/2})\|u\|_W, \\ \|u - u_h\|_V \leq cH^k\|u\|_W. \end{cases} \quad (2.16)$$

Proof. Voir [Gue01b].

Remark 2.12. Note that the error estimate in the V-norm is uniform with respect to ϵ. The uniformity is an improvement with respect to the Galerkin Least Square method.

2.5 Two-level $\mathbb{P}_1$ and $\mathbb{P}_2$ interpolation

In this section we describe two finite element settings that satisfy the hypotheses of the subgrid viscosity technique presented above. For the sake of simplicity we assume that Ω is a polyhedron in $\mathbb{R}^d$ and $\mathcal{T}_H$ is a regular triangulation of Ω composed of affine simplices (K_H). The reference simplex is denoted by $\hat{K}$ and $T_{K_H} : \hat{K} \longrightarrow K_H$ is the affine mapping that maps $\hat{K}$ onto K_H.

2.5.1 Definitions and preliminaries

To consider at once every linear first order differential operator, we introduce a family of d functions $(A^k)_{k=1,d}$ with values in the space of real matrices of order $m \times m$ where $m > 0$.; i.e., $A^k : \Omega \longrightarrow \mathcal{M}_m(\mathbb{R})$. We define the matrix field $\beta = (A^1, \ldots, A^d)$, and for a smooth function $u : \Omega \longrightarrow \mathbb{R}^m$ we denote by $\beta \cdot \nabla u$ the function $\beta \cdot \nabla u : \Omega \longrightarrow \mathbb{R}^m$ s.t.

$$1 \leq i \leq m, \qquad (\beta \cdot \nabla u)_i = \sum_{k=1}^{d} \sum_{j=1}^{m} A_{ij}^k \frac{\partial u_j}{\partial x_k}. \quad (2.17)$$

For a smooth function $v : \Omega \longrightarrow \mathbb{R}^m$, we set $\|v\|_{0,\Omega} = (\sum_{i=1}^{m} \|v_i\|_{0,\Omega}^2)^{1/2}$, and we define $v \cdot (\beta \cdot \nabla u) = \sum_{i=1}^{m} v_i(\beta \cdot \nabla u)_i$. We introduce also the semi-norm

$$|u|_{1,\beta,\Omega} = \left[\int_\Omega (\beta \cdot \nabla u) \cdot (\beta \cdot \nabla u) \right]^{1/2}.$$

2.5.2 Two-level $\mathbb{P}_1$ interpolation

We restrict ourselves to 2D, but all that is said can be generalized to 3D. Let us define first X_H by

$$X_H = \{v_H \in H^1(\Omega)^m;\ v_{H|K_H} \in \mathbb{P}_1(K_H)^m, \\ \forall K_H \in \mathcal{T}_H\}. \tag{2.18}$$

From each triangle $K_H \in \mathcal{T}_H$, we create 4 new triangles by connecting the middles of the 3 edges of K_H. Let us set $h = H/2$ and denote by $\mathcal{T}_h$ the resulting new triangulation. For each macro-triangle K_H, we define $\mathbb{P}$ as being the space of the functions that are continuous on K_H, vanish at the three vertices of K_H, and are piecewise $\mathbb{P}_1$ on each sub-triangle of K_H. We define

$$X_h^H = \{v_h^H \in H^1(\Omega)^m \mid v_{h|K_H}^H \in \mathbb{P}^m, \\ \forall K_H \in \mathcal{T}_H\}. \tag{2.19}$$

By setting $X_h = X_H \oplus X_h^H$, it is clear that we can characterize X_h by

$$X_h = \{v_h \in H^1(\Omega)^m \mid v_{h|K_h} \in \mathbb{P}_1(K_h)^m, \\ \forall K_h \in \mathcal{T}_h\}. \tag{2.20}$$

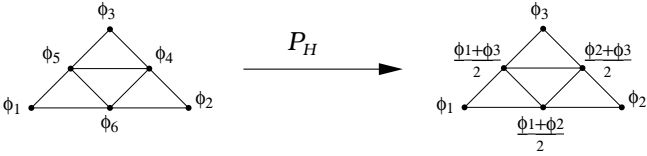The couple (X_H, X_h) is referred to as the two-level $\mathbb{P}_1$ setting.

Fig. 1. Definition of P_H for the two-level $\mathbb{P}_1$ setting.

On figure 1 we show a schematic representation of the action of the filter $P_H : X_h \longrightarrow X_H$ on a macro-element K_H of $\mathcal{T}_H$.

2.5.3 Two-level $\mathbb{P}_2$ interpolation

Let us build now a two-level $\mathbb{P}_2$ setting. Once more, we set $h = H/2$ and we denote by $\mathcal{T}_h$ the triangulation obtained by dividing each macro-triangle of $\mathcal{T}_H$ into 4 sub-triangles. For each triangle K_h, we denote by ψ_1, ψ_2, ψ_3 the three nodal $\mathbb{P}_2$ functions associated with the middle of the three edges of K_h. We set

$$X_H = \{v_H \in H^1(\Omega)^m;\ v_{H|K_H} \in \mathbb{P}_2(K_H)^m, \\ \forall K_H \in \mathcal{T}_H\}. \tag{2.21}$$

and we define the space of the subgrid scales as follows

$$X_h^H = \{v_h^H \in H^1(\Omega)^m; \\ v_{h|K_h}^H \in \text{vect}(\psi_1, \psi_2, \psi_3)^m,\ \forall K_h \in \mathcal{T}_h\}. \tag{2.22}$$

The space $X_h = X_H \oplus X_h^H$ is characterized by

$$X_h = \{v_h \in H^1(\Omega)^m;\ v_{h|K_h} \in \mathbb{P}_2(K_h)^m, \\ \forall K_h \in \mathcal{T}_h\}. \tag{2.23}$$

The couple (X_H, X_h) is called the two-level $\mathbb{P}_2$ setting.

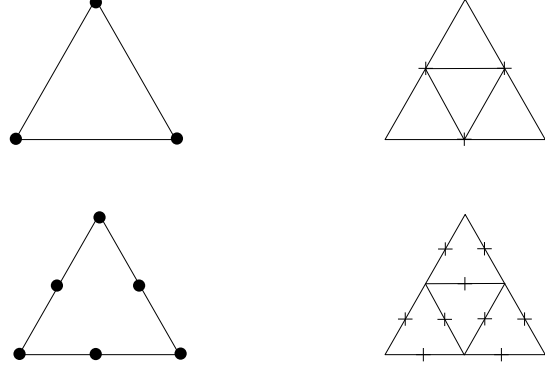The two interpolation settings described above are shown in figure 2.

Fig. 2. Two examples of hierarchical finite elements. Resolved scales spaces are on the left and subgrid scales on the right. From top to bottom: two-level $\mathbb{P}_1$; two-level $\mathbb{P}_2$.

2.5.4 The inf-sup condition

For the interpolation settings considered above, the decomposition $X_h = X_H \oplus X_h^H$ is L^2-stable. Furthermore, we have the following result.

Lemma 2.1. *If β is piecewise constant on each simplex K_H of $\mathcal{T}_H$, there is a constant $c_\beta > 0$, independent of (H, h), s.t. fo rall $u_H \in X_H$,*

$$\sup_{v_h \in X_h} \frac{\int_\Omega v_h \cdot (\beta \cdot \nabla u_H)}{\|v_h\|_{0,\Omega}} \geq c_\beta |u_H|_{1,\beta,\Omega}. \tag{2.24}$$

Corollary 2.1. *If β is in $C^1(\overline{\Omega}; \mathcal{M}_m(\mathbb{R})^d)$, there are two constants $c_\beta > 0$ and $c_\delta \geq 0$, both independent of (H, h), s.t. for all $u_H \in X_H$,*

$$\sup_{v_h \in X_h} \frac{\int_\Omega (\beta \cdot \nabla u_H) v_h}{\|v_h\|_{0,\Omega}} \geq c_\beta |u_H|_{1,\beta,\Omega} - c_\delta \|u_H\|_{0,\Omega}. \tag{2.25}$$

Proof. The reader is referred to [Gue99a] and [Gue99b] for the technical details.

Remark 2.13. The stabilizing properties of bubble functions for advection/diffusion problems have been put in evidence in [BBF+92]. Theoretical justifications can be found in [BBF93] and [BFHR97]. The importance of the inf-sup inequality (2.2) for problems like (1.1) does not seem to be well known by numericists.

2.6 Some examples

We show now that for the two problems considered in sections 1.2 and 1.3, the hypotheses (2.9), (2.10), and (2.11) are satisfied.

2.6.1 The advection/reaction problem

The advection/reaction problem of section 1.2 can be reformulated within our abstract framework by setting $m = 1$ and
$$A_{11}^k = \beta_k.$$
Let us define $a_0(u,v) = (\mu u, v)_{0,\Omega}, a_1(u,v) = (\beta \cdot \nabla u, v)_{0,\Omega}$ and $|u|_V = |u|_{1,\beta,\Omega}$. The hypothesis (2.9) is a simple consequence of the relation $a_s(u,u) \geq \mu_0 \|u\|_{0,\Omega}^2$ together with the definition of the semi-norm $|\cdot|_V$. The hypothesis (2.10) is a consequence of corollary 2.1 together with the $L^2(\Omega)$-coercivity of a_s. By setting
$$b(v_h^H, w_h^H) = c_b H(\nabla v_h^H, \nabla w_h^H)_{0,\Omega} \quad \text{and}$$
$$|v_h^H|_b = |v_h^H|_{1,\Omega},$$
the hypothesis (2.11) is obviously satisfied.

2.6.2 Le Darcy problem

Let us reformulate the Darcy problem considered in section 1.3 within our abstract framework. Let us set $m = d + 1$ and
$$\begin{aligned}
A_{ij}^k &= 0, && \text{if } 1 \leq i \leq m-1, 1 \leq j \leq m-1,\\
A_{ij}^k &= \delta_{i,k}, && \text{if } 1 \leq i \leq m-1, j = m,\\
A_{ij}^k &= \delta_{j,k}, && \text{if } i = m, \quad 1 \leq j \leq m-1,\\
A_{ij}^k &= 0, && \text{if } i = m, \quad j = m,
\end{aligned}$$
where $\delta_{i,k}$ is the Kronecker symbol. Define
$$a_0((u,p),(v,q)) = (K^{-1} \cdot u, v)_{0,\Omega}$$
$$a_1((u,p),(v,q)) = (\beta \cdot \nabla(u,p),(v,q))_{0,\Omega}.$$
It is clear that given definition (2.17), we have
$$a_1((u,p),(v,q)) = (q, \nabla \cdot u)_{0,\Omega} + (\nabla p, v)_{0,\Omega}.$$
Let us define $|(u,p)|_V = |(u,p)|_{1,\beta,\Omega}$. A simple calculation shows that $|(u,p)|_V = (\|\nabla \cdot u\|_{0,\Omega}^2 + \|\nabla p\|_{0,\Omega}^2)^{1/2}$. The hypothesis (2.9) is a simple consequence of the relation $a_s((u,p),(u,p)) = a_0((u,p),(u,p)) \geq \alpha' \|u\|_{0,\Omega}^2$ together with the definition of the semi-norm $|\cdot|_V$ and the Poincaré inequality. Since the matrix field β is constant on Ω, the hypothesis (2.10) is a consequence of lemma 2.1. By setting
$$\begin{aligned}
b((v_h^H, q_h^H),(w_h^H, r_h^H)) = c_b H((\nabla v_h^H, \nabla w_h^H)_{0,\Omega}\\
+ (\nabla q_h^H, \nabla r_h^H)_{0,\Omega})
\end{aligned}$$
and $|(v_h^H, q_h^H)|_b = (|v_h^H|_{1,\Omega}^2 + |q_h^H|_{1,\Omega}^2)^{1/2}$, the hypothesis (2.11) is obviously satisfied.

2.7 Numerical illustrations

2.7.1 Example 1: an advection problem

Let us consider the following problem.
$$\begin{cases} \partial_y u = \frac{1}{2\epsilon}(1 - (\tanh(\frac{y-0.5}{\epsilon}))^2) & \text{in } \Omega =]0,1[^2,\\ u_{|y=0} = 0, \end{cases} \quad (2.26)$$
where $u = \frac{1}{2}(\tanh(\frac{y-0.5}{\epsilon}) + 1)$ is the exact solution. We make numerical tests with $\epsilon = 0.04$. We use two-level $\mathbb{P}_1$ and $\mathbb{P}_2$ finite elements on a mesh $\mathcal{T}_h$ composed of 952 triangles and 517 vertices; i.e. $h \approx 1/20$. The bilinear form b_h is defined by
$$b_h(v_h^H, w_h^H) = c_b \sum_{K_h \in \mathcal{T}_h} \text{mes}(K_h)^{1/2} \int_{K_h} \nabla v_h^H \cdot \nabla w_h^H. \quad (2.27)$$
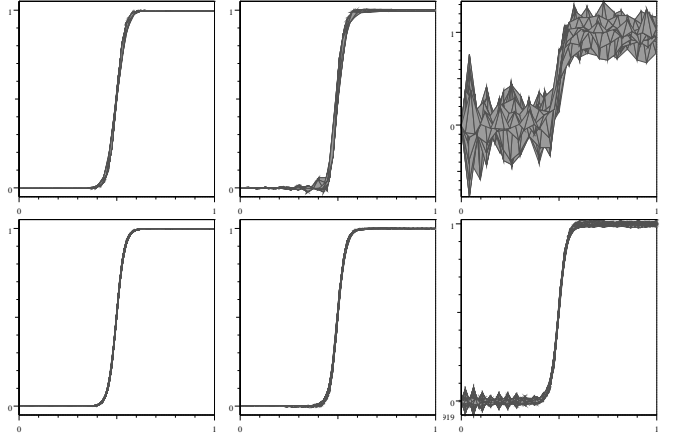We use $c_b = 1$. The results are shown in figure 3. The

Fig. 3. Problem (2.26). Projection in plane $x = 0$ of solution. Top; $\mathbb{P}_1$ solution. Bottom; $\mathbb{P}_2$ solution. From left to right; Lagrange interpolate of exact solution, stabilized solution, Galerkin solution.

projections in the plane $x = 0$ of the graphs of the $\mathbb{P}_1$ and $\mathbb{P}_2$ interpolates of the exact solution are shown on the left of the figure. The two-level $\mathbb{P}_1$ and $\mathbb{P}_2$ solutions are in the center; the Galerkin $\mathbb{P}_1$ and $\mathbb{P}_2$ solutions are on the right. The stabilizing effects of the subgrid viscosity method are clearly illustrated by this example.

2.7.2 Example 2 : boundary layer problem

We illustrate now the method on an advection/diffusion problem.
$$\begin{cases} \partial_y u - \nu \nabla^2 u = 0 & \text{in } \Omega =]0,1[^2,\\ u_{|y=0} = u_{|y=1} = 0, & \partial_x u_{|x=0} = \partial_x u_{|x=1} = 0, \end{cases} \quad (2.28)$$
where $u = (\exp(y/\nu) - 1)/(\exp(1/\nu) - 1)$ is the exact solution. We take $\nu = 0.002$ in the numerical tests. We
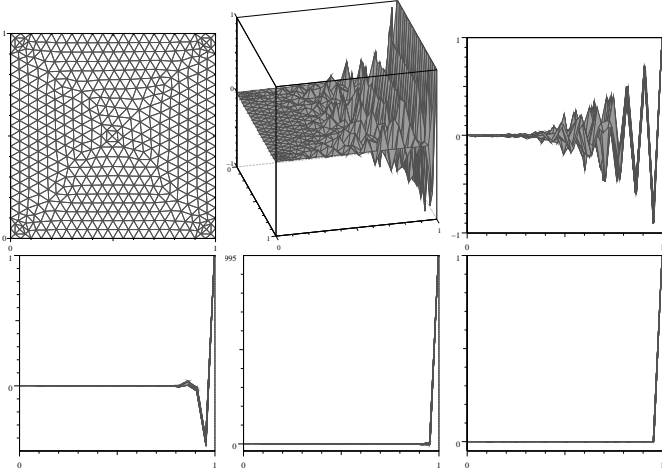
Fig. 4. Problem (2.28). Top left, mesh $\mathcal{T}_h$; top center, graph of $\mathbb{P}_1$ Galerkin solution; top right, projection in plane $x = 0$ of graph of Galerkin solution; bottom left, stabilized solution; bottom center, stabilized solution with shock capturing term; bottom right, $\mathbb{P}_1$ interpolate of exact solution.

use the same mesh as in the previous example and we approximate the solution by means of two-level $\mathbb{P}_1$ finite elements. The bilinear form b_h is the same as in (2.27). In figure 4 we show: the mesh (top left); the graph of the $\mathbb{P}_1$ Galerkin solution (top center); the projection in plane $x = 0$ of the Galerkin solution (top right), note the spurious oscillations spreading throughout the computation domain; the projection of the graph of the $\mathbb{P}_1$ interpolate of the solution (bottom right).

The stabilized solution is shown at the bottom left of figure 4. Note that all the spurious oscillations have disappeared except in the vicinity of the boundary layer where the slope of the solution is large. These residual oscillations are due to the Gibbs phenomenon. This phenomenon is well-known to numericists who work on non-linear conservation laws developing discontinuities. It is the manifestation of a far-reaching theorem in analysis that states that truncated Fourier series of a given function does not converge uniformly to the function in question unless the function is very smooth (continuity is not enough), see Rudin [Rud87, p. 97–98] for more details. A simple trick to eliminate this unwelcome oscillations consists of adding strong dissipation in the region of space where the solution is rough. Of course, one does not know a priori where the solution is rough, but one may expect that in this region the quantity $\nabla u_h^H = \nabla(u_h - P_H u_h)$ is of the same order as ∇u_h. Indeed, it is easy to show that, if u is a smooth function, denoting by $\mathcal{I}_h u$ the Lagrange interpolate of u, the quantity $\|\mathcal{I}_h u - P_H \mathcal{I}_h u\|_{0,\Omega}$ is of order h^{k+1} and $|\mathcal{I}_h u - P_H \mathcal{I}_h u|_{1,\Omega}$ is of order h^k. Hence, we are led to introduce the following nonlinear form:

$$c_h(u_h^H, v_h, w_h) = \qquad (2.29)$$

$$c_{sc} \sum_{K_H \in \mathcal{T}_H} \text{meas}(K_H)^{1/2} \frac{\|\nabla u_h^H\|_{0,K_H}}{\|\nabla u_h\|_{0,K_H}} \int_{K_H} (\nabla v_h \cdot \nabla w_h).$$

This form is called the shock capturing form. The modified problem we consider is the following.

$$\begin{cases} \text{Find } u_h \in X_h \text{ s.t. } \forall v_h \in X_h \\ a(u_h, v_h) + b_h(u_h^H, v_h^H) + c_h(u_h^H, u_h, v_h) = (f, v_h). \end{cases}$$

Since the nonlinearity is small, this problem can be solved by means of a very crude fixed point algorithm. The solution is shown in figure 4 at the bottom center location. The efficiency of the shock capturing form is evident. The boundary layer is captured within one element with $c_{sc} = 0.1$.

3 Evolution problem with non coercivity

In this section we show how the subgrid viscosity technique can be extended to treat time-dependent problems with no coercivity.

3.1 The model problem

The goal of this section is to introduce a general framework for non-coercive time-dependent problems. Let L be a separable Hilbert space and $A : D(A) \subset L \longrightarrow L$ be a linear operator.

Definition 3.1. *We say that A is monotone iff*

$$\forall v \in D(A), \qquad (Av, v)_L \geq 0, \qquad (3.1)$$

and A is maximal iff

$$\forall f \in L, \ \exists v \in D(A), \qquad v + Av = f. \qquad (3.2)$$

Lemma 3.1. *If $A : D(A) \subset L \longrightarrow L$ is maximal and monotone, then*
(i) $D(A)$ is dense in L.
(ii) The graph of A is closed.
(iii) For all $\lambda > 0$, $I + \lambda A : D(A) \subset L \longrightarrow L$ is bijective and $(I + \lambda A)^{-1}\|_{\mathcal{L}(L,L)} \leq 1$.

Proof. See Brezis [Bre91, p. 101], Showalter [Sho96, p. 22] or Yosida [Yos80, p. 246]. □

The major result of this section is the following.

Theorem 3.1 (Hille–Yosida). *For all $f \in \mathcal{C}^1([0, +\infty[; L)$ and all $u_0 \in D(A)$, the problem*

$$\begin{cases} \text{Find } u \in \mathcal{C}^1([0, +\infty[; L) \cap \mathcal{C}^0([0, +\infty[; D(A)) \text{ s.t.} \\ u_{|t=0} = u_0, \\ d_t u + Au = f, \end{cases} \quad (3.3)$$

has a unique solution and

$$\begin{cases} \|u\|_{\mathcal{C}^0([0,T];L)} \leq c\left(\|u_0\|_L + T\|f\|_{\mathcal{C}^0([0,T];L)}\right), \\ \|u\|_{\mathcal{C}^1([0,T];L)} + \|u\|_{\mathcal{C}^0([0,T];V)} \\ \qquad \leq c\left(\|u_0\|_V + T\|f\|_{\mathcal{C}^1([0,T];L)}\right). \end{cases} \quad (3.4)$$

Proof. See [Bre91, p. 110] or Yosida [Yos80, p. 248]. □

To reformulate problem (3.3), we introduce the bilinear form a such that $a(u,v) = (Au, v)_L$ for all $u \in D(A)$ and $v \in L$. We set $V = D(A)$ and we equip V with the graph norm: $\|v\|_V = (\|v\|_L^2 + \|Av\|_L^2)^{1/2}$. Since the graph of A is closed, lemma 3.1 implies that V is a Banach space. Hence, the bilinear form $a : V \times L \longrightarrow \mathbb{R}$ is continuous. Furthermore, when equipped with the scalar product $(u,v)_L + (Au, Av)_L$, V is a Hilbert space. Since $D(A) = V$ is dense in L (lemma 3.1), we are in the classical situation $V \subset L \equiv L' \subset V'$.

We reformulate problem (3.3) as follows. For $f \in \mathcal{C}^1([0, +\infty[; L)$ and $u_0 \in V$,

$$\begin{cases} \text{Find } u \text{ in } \mathcal{C}^1([0, +\infty[; L) \cap \mathcal{C}^0([0, +\infty[; V) \text{ s.t.} \\ (u(0), v) = (u_0, v), \quad \forall v \in L, \\ (d_t u, v)_L + a(u, v) = (f, v)_L, \quad \forall v \in L, \ \forall t \geq 0. \end{cases} \quad (3.5)$$

Remark 3.1. This problem is strictly equivalent to the original problem (3.3). The Hille-Yosida theorem guarantees that it is well-posed.

Remark 3.2. The reader can very that the advection reaction operator and the Darcy operator introduced in sections 1.2 and 1.3 are maximal and monotone.

Let us introduce the semi-norm $|v|_V = \|Av\|_L$.

Proposition 3.1. *Let $A \in \mathcal{L}(V; L)$ be a monotone operator. The following two properties are equivalent.*
(i) A is maximal.
(ii) There are two constants $c_1 > 0$, $c_2 \geq 0$ such that

$$\forall u \in V, \qquad \sup_{v \in L} \frac{a(u,v)}{\|v\|_L} = c_1 |u|_V - c_2 \|u\|_L. \quad (3.6)$$

Remark 3.3. In general, if the bilinear form a is not coercive, when using the Galerkin technique to build an approximate solution to problem (3.5), the inequality (3.6) is not satisfied uniformly with respect to the mesh size.

To build an optimal approximate solution to problem (3.5), one possible approach consists in generalizing the Galerkin Least Square technique. This choice implies that no difference is made between space and time, the consequence being that a discontinuous Galerkin approximation of time must be done. The reader interested in this approach is referred to [CKS00], [LR74], [Joh87] or [JNP84].

The other approach that we shall develop herein consists in using the subgrid viscosity technique.

3.2 The subgrid viscosity technique

Let us recall the discrete setting introduced in section 2.1. Let $X_H \subset X_h \subset V$ be two sequences of finite dimensional spaces satifying (2.1).

We assume that a discrete version of (3.6) is satisfied. More precisely, there are $c_a > 0$ and $c_\delta \geq 0$, independent of (H, h) such that for all $v_h \in X_h$,

$$\sup_{\phi_h \in X_h} \frac{a(v_H, \phi_h)}{\|\phi_h\|_L} \geq c_a |v_H|_V - c_\delta \|v_h\|_L. \quad (3.7)$$

Furthermore, we assume that the hypotheses (2.3), (2.4), and (2.11) hold true. We refer to sections 2.1 and 2.5 for a discussion on these hypotheses and examples of admissible finite element couples (X_h, X_H).

Let us assume that $u_0 \in W$ such that u_0 can be approximated by $I_H u_0$. The discrete problem we consider reads:

$$\begin{cases} \text{Find } u_h \in \mathcal{C}^1([0, +\infty[; X_h) \text{ s.t. } \forall v_h \in X_h \\ (d_t u_h, v_h)_L + a(u_h, v_h) + b_h(u_h^H, v_h^H) = (f, v_h), \\ u_{h|t=0} = I_H u_0. \end{cases} \quad (3.8)$$

This problem has a unique solution, for it is a system of linear ordinary differential equations.

The major convergence result of this section is the following.

Theorem 3.2. *Under hypotheses (2.1), (3.7), (2.3), (2.4), and (2.11), if u is in $\mathcal{C}^2([0, T]; W)$, then u_h satisfies the following error estimates.*

$$\|u - u_h\|_{\mathcal{C}^0([0,T];L)}$$
$$+ \left[\int_0^T a(u - u_h, u - u_h) \right]^{1/2} \leq c_1 H^{k+1/2}, \quad (3.9)$$

$$\left[\frac{1}{T} \int_0^T \|u - u_h\|_V^2 \right]^{1/2} \leq c_2 H^k, \quad (3.10)$$

where constants c_1 and c_2 are bound from above as follows.

$$c_1 \leq c \left[H + T (1 + T) \right]^{1/2} \|u\|_{\mathcal{C}^2([0,T];W)},$$
$$c_2 \leq c \left[1 + T \right] \|u\|_{\mathcal{C}^2([0,T];W)}.$$

Remark 3.4. Note that the norm used in the error estimates are the same as those of the stability estimates (3.4). The estimate (3.10) is optimal in the graph norm. The estimate (3.9) is the same as that obtained by the Discontinuous Galerkin technique [JNP84].

Remark 3.5. Note when T is large $c_1 = \mathcal{O}(T)$ and $c_2 = \mathcal{O}(T)$; that is, in the most unfavorable case the error increase linearly with T.

3.3 A singular perturbation problem

The technique developed above is tailored for first order differential operators. In practice, we have to deal with situations where $B = A + \epsilon D$, A is a first order differential operator and D is a coercive second order differential operator. From the mathematical point of view, the coercivity of D implies that the evolution equation is parabolic. If ϵ is $\mathcal{O}(1)$, the problem falls within the framework of parabolic equations whose approximation by the Galerkin technique is optimal. On the other hand, if ϵ is small, the coercivity is not strong enough to guarantee that the Galerkin approximation is satisfactory, for in first approximation $B \approx A$. We show now that the subgrid viscosity technique can easily be extended to treat this situation.

Let us retain the notation introduced above. In addition to the two Hilbert spaces already introduced, L and $D(A) = V$, we introduce a new Hilbert space X with dense and continuous embedding in V. We introduce also a bilinear form $d \in \mathcal{L}(X \times X; \mathbb{R})$, and we assume that there is a semi-norm $|\cdot|_X$ in X so that $d(u,v) \leq c_d |u|_X |v|_X$ for all u, v in X. In practice D can be a degenerate elliptic operator. We assume that $a + d$ is coercive with respect to the semi-norm $|\cdot|_X$, i.e.

$$\forall v \in X, \qquad |v|_X^2 \leq a(v,v) + d(v,v). \qquad (3.11)$$

Let us introduce the following space

$$W(X) = \{v \in L^2(0,+\infty; X);\ d_t v \in L^2(0,+\infty; X')\}$$

We consider now the following problem: for $u_0 \in X$ and $f \in C^1([0,+\infty[; L)$,

$$\begin{cases} \text{Find } u \text{ in } W(X) \text{ s.t. } \forall v \in X,\ \forall t \geq 0 \\ (d_t u, v)_L + a(u,v) + \epsilon d(u,v) = (f,v)_L, \qquad (3.12) \\ (u(0), v) = (u_0, v), \quad \forall v \in L, \end{cases}$$

where ϵ is a positive real number which may possibly be zero. We assume that the problem is normalized so that $\epsilon \leq 1$. Furthermore, we assume that there is $c > 0$ so that $\|v\|_X \leq c(\|v\|_L + |v|_X)$. The consequence of this hypothesis is that problem (3.12) is parabolic in Lions' sense [LM68, p. 253] and has a unique solution.

Now we use the discrete setting of §3.2 to build an approximate solution to problem (3.12). Let us introduce two sequences of finite dimensional spaces $X_H \subset X_h \subset X$ satisfying hypotheses (2.1), (3.7), (2.3), (2.4), and (2.11). Furthermore, we assume that the following inverse inequality holds

$$|v_h|_X \leq c H^{-1} \|v_h\|_L. \qquad (3.13)$$

We assume that $u_0 \in W$ so that $I_H u_0$ is a good approximation to u_0. The discrete problem that we consider reads

$$\begin{cases} \text{Find } u_h \text{ in } C^1([0,+\infty[; X_h) \text{ s.t. } \forall v_h \in X_h \\ (d_t u_h, v_h)_L + a(u_h, v_h) + \epsilon d(u_h, v_h) \\ \qquad + b_h(u_h^H, v_h^H) = (f, v_h), \\ u_{h|t=0} = I_H u_0. \end{cases} \qquad (3.14)$$

Problem (3.14) is well-posed since it is a linear system of ordinary differential equations.

Theorem 3.3. *If u is in $C^2([0,T]; W)$, then u_h, solution to (3.14), satisfies*

$$\|u - u_h\|_{C^0([0,T];L)} + \left[\int_0^T a_s(u-u_h, u-u_h)\right]^{1/2}$$
$$+ \epsilon^{1/2} \|u - u_h\|_{L^2([0,T];X)} \qquad (3.15)$$
$$\leq c_1(T,u)\left[H^{k+1/2} + \epsilon^{1/2} H^k\right],$$

$$\left[\frac{1}{T}\int_0^T \|u - u_h\|_V^2\right]^{1/2} \leq c_2(T,u) H^k, \qquad (3.16)$$

where constants c_1 and c_2 are bounded from above as follows

$$c_1 \leq c\left[H + T(1+T)\right]^{1/2} \|u\|_{C^2([0,T];W)},$$
$$c_2 \leq c\left[1 + T\right]\|u\|_{C^2([0,T];W)}.$$

Proof. See [Gue01a]. $\qquad \blacksquare$

3.4 Some numerical examples

We evaluate the performance of the method by testing it on problems of increasing difficulties.

3.4.1 Example 1 : Advection in 1D

Let us first make convergence tests in space on the following $1D$ linear advection problem.

$$\begin{cases} u_{|t=0} = \sin(2\pi x^\alpha), \\ \partial_t u + \partial_x u = 0, \qquad \text{in } \Omega =]0,1[, \qquad (3.17) \\ \text{Periodic boundary condition.} \end{cases}$$

The exact solution is $u = \sin(2\pi(x-t)^\alpha)$. The problem falls within the framework developed above when setting

$L = L^2(\Omega)$,
$V = \{v \in L^2(\Omega) \mid \partial_x v \in L^2(\Omega),\ v_{|x=0} = v_{|x=1}\}$,
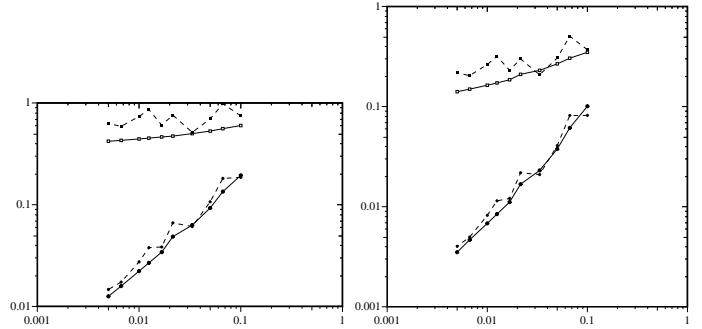$A = \partial_x$.

Fig. 5. Convergence tests; L^2 and H^1 norms with respect to h. Solid line : stabilized $\mathbb{P}_1$ solution; discontinuous line: Galerkin $\mathbb{P}_1$ solution. Left: $u_0 = \sin(2\pi x^{0.6})$; right: $u_0 = \sin(2\pi x^{0.8})$).

We approximate the solution by means of the two-level $\mathbb{P}_1$ finite elements defined in §2.5. By setting $\beta = (1,0)$, lemma 2.1 guarantees that the discrete inf-sup condition (3.7) is satisfied. We define b_h as in (2.27). In our tests $c_b = 0.1$. The family of meshes considered is regular, but to avoid super-convergence phenomena each mesh is obtained by a random mapping of the uniform grid with the same number of nodes.

To approximate the time derivative, we use the second order BDF2 scheme. The time step δt is chosen small

enough to guarantee that the time error is much smaller than the space error, i.e., $\delta t = 10^{-3}$. The total integration time is $T = 1$; i.e., the solution has crossed the domain once.

Convergence tests with $\alpha = 0.6$ and $\alpha = 0.8$ are reported in figure 5. In both cases the solution is in $C^0([0, +\infty[; H^1(\Omega))$. We plot the L^2 and H^1 norms of the error as a function of h for the stabilized solution and the Galerkin solution. It is clear that the convergence properties of the stabilized solution are superior to that of the Galerkin solution.
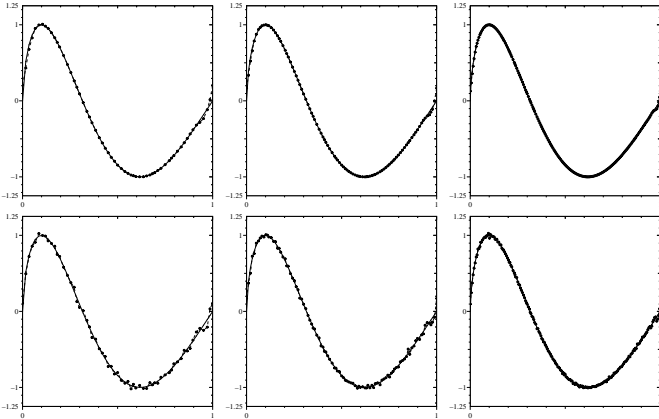

Fig. 6. Convergence tests with $u_0 = \sin(2\pi x^{0.6})$; top: stabilized $\mathbb{P}_1$ solution; bottom: Galerkin $\mathbb{P}_1$ solution; from left to right: $h = 1/60$, $h = 1/100$, $h = 1/200$.

To illustrate the convergence problems of the Galerkin approximation, we show in figure 6 the stabilized solution and the Galerkin solution on three different meshes: $h = 1/60$, $h = 100$ and $h = 1/200$. Note that for the three meshes considered, the Galerkin solution is polluted by spurious numerical oscillations spreading all over the domain, whereas the stabilized solution exhibits some very localized oscillations close to point where the first derivative is singular.

3.4.2 Example 2 : Advection in 2D

To illustrate further the performance of the subgrid stabilization technique, we solve problem (3.17) in 2D, $\Omega =]0, 1[^2$, with periodic boundary conditions and $\alpha = 0.6$. We use the $\mathbb{P}_1$ approximation on a mesh composed of 3728 triangles and 1945 nodes, i.e., $h \approx 1/40$. The solution at $T = 1$ is shown in figure 7. As in 1 dimension, the Galerkin solution oscillates throughout the domain, whereas the stabilized solution is smooth almost everywhere, except in the vicinity of the line where $\partial_x u$ is singular.

3.4.3 Example 3 : Advection with rough data

We now make a test with rough initial data on the 1D advection equation
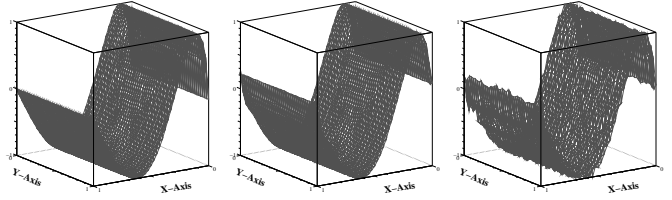
$$u_t + u_x = 0, \quad \text{in } \Omega =]-1, +1[,$$

Fig. 7. Advection problem (3.17) with $\alpha = 0.6$ in $\Omega =]0, 1[^2$. left: $\mathbb{P}_1$ interpolate of the exact solution; center: stabilized solution; right: Galerkin solution.

with periodic boundary conditions. The initial data proposed in [SO89] is

$$u_0(x) = \begin{cases} e^{-300(x+0.7)^2} & \text{if } |x + 0.7| \leq 0.25, \\ 1 & \text{if } |x + 0.1| \leq 0.2, \\ \left(1 - \left(\frac{x-0.6}{0.2}\right)^2\right)^{1/2} & \text{if } |x - 0.6| \leq 0.2, \\ 0 & \text{else.} \end{cases} \quad (3.18)$$

Like for time independent first order PDE's, the fact that the data u_0 is not in $D(A)$ triggers the Gibbs phenomenon. See example 2 of section 2.7 for details on this problem. To limit this phenomenon we introduce the nonlinear form (2.29). The approximate problem reads:

$$\begin{cases} \text{Find } u_h \text{ in } C^1([0, +\infty[; X_h) \text{ s.t. } \forall v_h \in X_h \\ (d_t u_h, v_h)_L + a(u_h, v_h) + \epsilon d(u_h, v_h) \\ \quad + b_h(u_h^H, v_h^H) + c_h(u_h^H, u_h, v_h) = (f, v_h), \\ u_{h|t=0} = I_H u_0, \end{cases} \quad (3.19)$$

We make tests with two-level $\mathbb{P}_1$ finite elements on three different grids composed of 50, 100, and 200 nodes respectively. We set $c_b = 0.05$ and $c_{sc} = 0.05$. We use BDF2 with $\delta t = 10^{-3}$ to march in time.

The two-level solution at $T = 4$ on the three considered meshes is plotted at the top of figure 8. The Galerkin solution is shown at the bottom of the figure. The Galerkin solution is of no use to engineers. It is clear that the stabilization is efficient and that the stabilized $\mathbb{P}_1$ solution converges to the exact solution satisfactorily.

3.4.4 Example 4 : A nonlinear degenerate parabolic problem

To test the capability of the proposed method to deal with degenerate parabolic problems, we consider a new class of convection-diffusion equations proposed in a series of papers by Kurganov and Rosenau [JGR99]. "The novel feature of these equations is that large amplitude solutions develop spontaneous discontinuities, while small solutions remain smooth at all times."

Let us consider the following problem in $\Omega =]-3, 3[$

$$\begin{cases} u_{|t=0} = \begin{cases} 1.2 & \text{if } -3 \leq x < 0, \\ -1.2 & \text{if } 0 < x \leq 3, \end{cases} \\ u(\pm 3, t) = \mp 1.2 & \text{for } 0 \leq t, \\ \partial_t u + \partial_x u^2 - \partial_x \left(\frac{\partial_x u}{\sqrt{1 + (\partial_x u)^2}}\right) = 0. \end{cases} \quad (3.20)$$
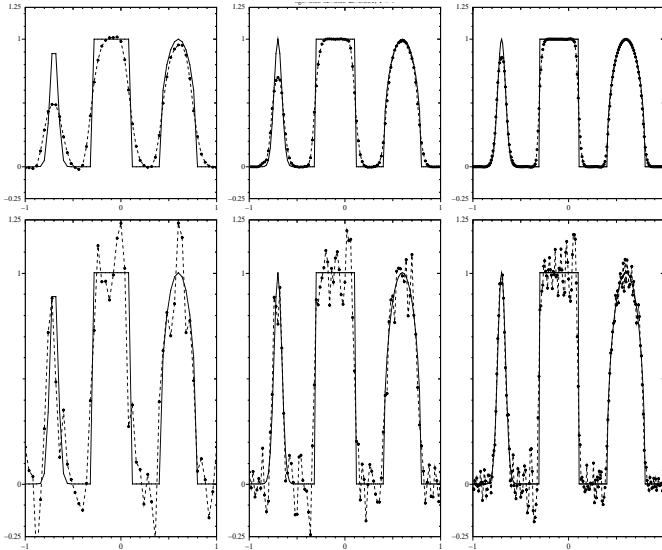
Fig. 8. 1D advection problem with rough initial data. Top: stabilized $\mathbb{P}_1$ solution; bottom: Galerkin $\mathbb{P}_1$ solution; From left to right: 50 nodes, 100 nodes, 200 nodes.
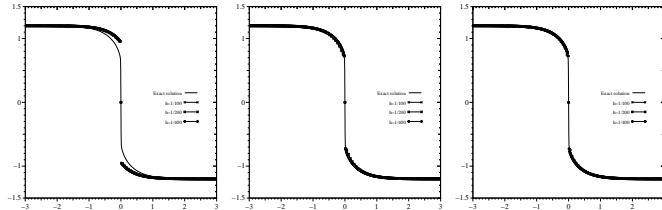

Fig. 9. Degenerate parabolic problem on three grids: $h = 6/100$, $h = 6/200$, and $h = 6/400$; left: $\mathbb{P}_1$ Galerkin solution; center: two-level $\mathbb{P}_1$ approximation with $c_b = c_{sc} = 0.2$, right: two-level $\mathbb{P}_1$ approximation with $c_b = 0.5$, $c_{sc} = 0.1$.

The problem is solved, up to time $T = 1.5$, by using formulation (3.19) with $\mathbb{P}_1$ finite elements on three grids: $h = 6/100$, $h = 6/200$, and $h = 6/400$. The results are shown in figure 9. Quite surprisingly, the Galerkin solution is not plagued by spurious oscillations but converges to a non-entropic solution. To illustrate the insensitivity of the method to variations on the stabilizing parameters, we make two sets of computations. In the first set we use $c_b = 0.2$, $c_{sc} = 0.2$ and in the other set we use $c_b = 0.5$, $c_{sc} = 0.1$. The results shown in figure 9 demonstrate that the stabilized solution converges and does not depend too much on the choice of the stabilizing parameters.

3.4.5 Example 5 : The compressible Navier–Stokes equations

To further illustrate the capability of the method we solve a compressible Navier–Stokes problem by Tenaud–Daru [VD00]. We consider a box $\Omega =]0, 1[^2$ filled with a viscous ideal gas. A diaphragm situated at $x = 1/2$ separates the box into two parts. The fluid is initially at rest and in two different thermodynamic states on each sides of the diaphragm. On the left we have $\rho_l = 120$

and $p_l = \rho_l/\gamma$, whereas on the right we have $\rho_r = 1.2$ and $p_r = \rho_r/\gamma$. The constant γ is set to 1.4. At $t = 0$ the diaphragm is removed. The shock moves to the right of the box, then reflects on the right side. When coming back to the left, the shock strongly interacts with the boundary layer that it created at the bottom of the box. The interaction produces a λ shock and a massive separation of the boundary layer.
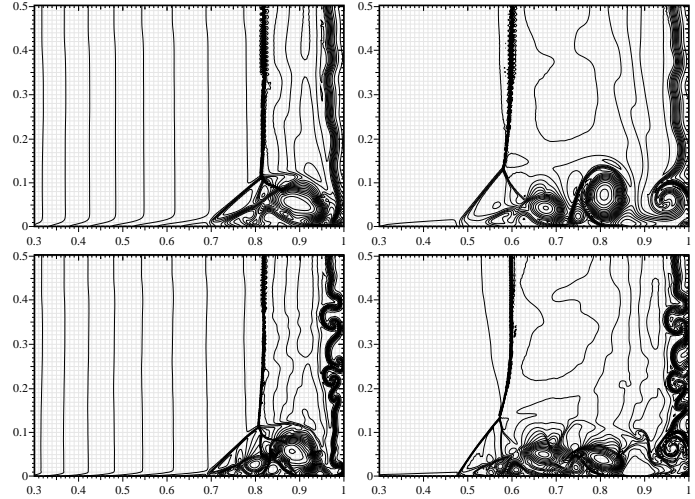

Fig. 10. Shock box problem; $\mathbb{P}_1$ approximation; density contour lines; Reynolds 200 (top); Reynolds 1000 (bottom); $t = 0.6$ (left); $t = 1$ (right).

The solution is assumed to be symmetric with respect to the axis $y = 1/2$; as a result, the computational domain is restricted to $\Omega =]0, 1[\times]0, 1/2[$. We use two-level $\mathbb{P}_1$ finite elements. Two Reynolds numbers are considered: $Re = 200$ and $Re = 1000$. The Prandtl number is set to 0.73. In figure 10 we show density contours for these two Reynolds numbers at times $T = 0.6$ and $T = 1$. The contour step is $\Delta\rho = 5$, and the contour lines are shown from $\rho = 10$ to $\rho = 120$. The solution shown here compares quite well with that reported in [VD00].

References

[AP96]   P. Azerad and G. Pousin. Inégalité de poincaré courbe pour le traitement variationnel de l'équation de transport. *C. R. Acad. Sci. Paris, Sér. I*, 322(8):721–727, 1996.

[BBF⁺92] F. Brezzi, M.O. Bristeau, L. Franca, M. Mallet, and G. Rogé. A relationship between stabilized finite element methods and the Galerkin method with bubble functions. *Comput. Methods Appl. Mech. Engrg.*, 96:117–129, 1992.

[BBF93] C. Baiocchi, F. Brezzi, and L.P. Franca. Virtual bubbles and Galerkin-Least-Square type methods (GaLS). *Comput. Methods Appl. Mech. Engrg.*, 105:125–141, 1993.

[BFHR97] F. Brezzi, L. Franca, T.J.R. Hughes, and A. Russo. $b = \int g$. *Comput. Methods Appl. Mech. Engrg.*, 145:329–364, 1997.

[Bre91]   H. Brezis. *Analyse fonctionnelle*. Masson, Paris, 1991.

[CKS00]   B. Cockburn, G.E. Karniadakis, and C.W. Shu. *Discontinuous Galerkin methods - theory, computation and applications*, volume 11 of *LNCSE*. Springer, 2000.

[Gue99a] J.-L. Guermond. Stabilisation par viscosité de sous-maille pour l'approximation de Galerkin des opérateurs linéaires monotones. *C. R. Acad. Sci. Paris, Sér. I*, 328:617–622, 1999.

[Gue99b] J.-L. Guermond. Stabilization of Galerkin approximations of transport equations by subgrid modeling. *Mod. Math. Anl. Numér. (M2AN)*, 33(6):1293–1316, 1999.

[Gue01a] J.-L. Guermond. Subgrid stabilization of Galerkin approximations of linear contraction semi-groups of class C^0 in Hilbert spaces. *Numerical Methods for Partial Differential Equations*, 17:1–25, 2001.

[Gue01b] J.-L. Guermond. Subgrid stabilization of Galerkin approximations of linear monotone operators. *IMA, J. Numer. Anal.*, 21:165–197, 2001.

[JGR99]   A. Kurganov J. Goodman and P. Rosenau. Breakdown in burgers-type equations with saturating dissipation fluxes. *Nonlinearity*, 12:247–268, 1999.

[JNP84]   C. Johnson, U. Nävert, and J. Pitkäranta. Finite element methods for linear hyperbolic equations. *Comput. Methods Appl. Mech. Engrg.*, 45:285–312, 1984.

[Joh87]   C. Johnson. *Numerical solution of partial differential equations by the finite element method*. Cambridge University Press, Cambridge, 1987.

[LM68]   J.-L. Lions and E. Magenes. *Problèmes aux limites non homogènes et applications*, volume 1. Dunod, Paris, 1968.

[LR74]   P. Lesaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In C. de Boors, editor, *Mathematical aspects of Finite Elements in Partial Differential Equations*, pages 89–123, Academic Press, 1974.

[Neč62]   J. Nečas. Sur une méthode pour résoudre les équations aux dérivées partielles de type elliptique, voisine de la variationnelle. *Ann. Scuola Norm. Sup. Pisa*, 16:305–326, 1962.

[Rud87]   W. Rudin. *Analyse réelle et complexe*. Masson, Paris, 4ème edition, 1987.

[Sho96]   R.E. Showalter. *Monotone operators in Banach spaces and nonlinear partial differential equations*, volume 49 of *MSM*. AMS, 1996.

[SO89]   C.W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes, ii. *J. Comput. Phys.*, 83:32–78, 1989.

[Tad89]   E. Tadmor. Convergence of spectral methods for nonlinear conservation laws. *SIAM J. Numer. Anal.*, 26(1):30–44, 1989.

[VD00]   C. Tenaud V. Daru. Evaluation of tvd high resolution schemes for the unsteady viscous shocked flows. *Computers and Fluids*, 2000.

[Yos80]   K. Yosida. *Functional analysis*. Springer-Verlag, 6ème edition, 1980.

[Zho97]   G. Zhou. How accurate is the streamline diffusion finite element method? *Math. Comp.*, 66:31–44, 1997.