Linear stabilization for first-order PDEs

Alexandre Ern[a], Jean-Luc Guermond[b]

[a] *Université Paris-Est, CERMICS (ENPC), 77455 Marne-la-Vallée cedex 2, France.*
[b] *Department of mathematics, Texas A&M University, College Station, TX 77843-3368, USA.*

Abstract

We analyze finite-element-based stabilization techniques for first-order PDEs within the framework of symmetric Friedrichs systems, including residual-based methods like Galerkin Least Squares (GaLS) and fluctuation-based methods like Continuous Interior Penalty (CIP), Local Projection Stabilization (LPS) and Subgrid Viscosity (SGV).

Keywords: First-order PDEs, Linear stabilization, Galerkin Least-Squares, Streamline Diffusion, Subgrid viscosity,
2010 MSC: 35F05, 35F15, 65N12, 65N30, 65J10.

1. Friedrichs' systems

The objective of this section is to present the theory of the symmetric positive systems of first-order linear PDEs. This theory has been developed in 1958 by Friedrichs [18] to study transonic flows. Friedrichs wanted to handle within a single functional framework PDEs that are partly elliptic and partly hyperbolic, and for this purpose he developed a formalism that goes beyond the traditional classification of PDEs into elliptic, parabolic, and hyperbolic types. Friedrichs' formalism is very powerful and encompasses several model problems. Important examples are the advection-reaction equation, the div-grad problem related to Darcy's equations, and the curl-curl problem related to Maxwell's equations. This theory is an important key to understand stabilization techniques for first-order PDEs. All the theoretical arguments are presented assuming that the functions are complex-valued.

1.1. Basic ideas and model problem

Let D be a strongly Lipschitz domain in $\mathbb{R}^d$. We consider functions defined over D with values in $\mathbb{C}^m$, $m \geq 1$. Let $\mathcal{B}, \mathcal{C} \in \mathbb{C}^{m\times m}$ be two Hermitian matrices, i.e., $\mathcal{B} = \mathcal{B}^{\mathsf{H}}$, $\mathcal{C} = \mathcal{C}^{\mathsf{H}}$, where $\mathcal{Z}^{\mathsf{H}}$ is the Hermitian transpose of $\mathcal{Z}$; we say that $\mathcal{B} \geq \mathcal{C}$ if and only if $X^{\mathsf{H}}\mathcal{B}X \geq X^{\mathsf{H}}\mathcal{C}X$ for all $X \in \mathbb{C}^m$.

Email addresses: alexandre.ern@enpc.fr (Alexandre Ern), guermond@math.tamu.edu (Jean-Luc Guermond)

Let $\mathcal{K}$, $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ be a family of $(d+1)$ fields on D with values in $\mathbb{C}^{m\times m}$. We assume that these fields satisfy the following key assumptions:

$$\text{Boundedness:} \quad \mathcal{K},\ \{\mathcal{A}^k\}_{k\in\{1:d\}},\ \text{and } \mathcal{X} \text{ are in } L^\infty(D;\mathbb{C}^{m\times m}), \tag{1a}$$

$$\text{Symmetry:} \quad \mathcal{A}^k = (\mathcal{A}^k)^{\mathsf{H}} \text{ for all } k\in\{1:d\}, \text{ a.e. in } D, \tag{1b}$$

$$\text{Positivity:} \quad \exists \mu_0>0 \text{ s.t. } \mathcal{K}+\mathcal{K}^{\mathsf{H}}-\mathcal{X}\geq 2\mu_0\mathbb{I}_m \text{ a.e. in } D. \tag{1c}$$

In (1c), $\mathbb{I}_m$ denotes the identity matrix in $\mathbb{C}^{m\times m}$ and $\mathcal{X}:=\sum_{k=1}^d \partial_k\mathcal{A}^k$ where $\partial_k := \frac{\partial}{\partial x_k}$. Note that $\mathcal{X}=\mathcal{X}^{\mathsf{H}}$ owing to (1b). We now define two differential operators A and A_1 such that

$$Av := \mathcal{K}v + A_1 v, \qquad A_1 v := \sum_{k\in\{1:d\}} \mathcal{A}^k\partial_k v, \qquad \forall v\in C^1(\overline{D};\mathbb{C}^m). \tag{2}$$

In what follows, we assume that the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ have a bounded trace at the boundary ∂D, and we introduce the boundary field $\mathcal{N}\in L^\infty(\partial D;\mathbb{C}^{m\times m})$ such that $\mathcal{N}:=\sum_{k\in\{1:d\}} n_k\mathcal{A}^k_{|\partial D}$, where $(n_k)_{k\in\{1:d\}}$ are the Cartesian components of the outward unit normal $\boldsymbol{n}$. Note that $\mathcal{N}=\mathcal{N}^{\mathsf{H}}$ owing to (1b).

Let $L:=L^2(D;\mathbb{C}^m)$ and let us denote $(f,g)_L := \int_D g^{\mathsf{H}} f\,\mathrm{d}x$ for any $f,g\in L$; note that $(f,g)_L = \overline{(g,f)_L}$. Similarly we introduce $L^\partial := L^2(\partial D;\mathbb{C}^m)$ with the inner product $(f,g)_{L^\partial} := \int_{\partial D} g^{\mathsf{H}} f\,\mathrm{d}s$. Integration by parts using the (Hermitian) inner product in L is a key tool in the analysis of Friedrichs' systems. To formalize this idea we define the formal adjoint $\tilde{A}$ of A such that

$$\tilde{A}v := (\mathcal{K}^{\mathsf{H}}-\mathcal{X})v - A_1 v = (\mathcal{K}+\mathcal{K}^{\mathsf{H}}-\mathcal{X})v - Av, \qquad \forall v\in C^1(\overline{D};\mathbb{C}^m). \tag{3}$$

Lemma 1.1 (Integration by parts)**.** *The following holds for all $v,w\in C^1(\overline{D};\mathbb{C}^m)$:*

$$(Av,w)_L = (v,\tilde{A}w)_L + (\mathcal{N}v,w)_{L^\partial}, \tag{4}$$

$$\Re\left((Av,v)_L\right) \geq \mu_0\|v\|_L^2 + \frac{1}{2}(\mathcal{N}v,v)_{L^\partial}. \tag{5}$$

The lower bound (5) says that the sesquilinear form $(Av,w)_L$ is L-coercive up to a boundary term. The key idea of Friedrichs is to enforce a suitable boundary condition to gain positivity on the boundary term. This is done by assuming that there exists another boundary field $\mathcal{M}\in L^\infty(\partial D;\mathbb{C}^{m\times m})$ satisfying the following two algebraic properties a.e. on ∂D:

$$\mathcal{M} \text{ is non-negative: } \Re(\xi^{\mathsf{H}}\mathcal{M}\xi)\geq 0 \text{ for all } \xi\in\mathbb{C}^m, \tag{6a}$$

$$\ker(\mathcal{M}-\mathcal{N})+\ker(\mathcal{M}+\mathcal{N})=\mathbb{C}^m. \tag{6b}$$

Since any function v satisfying $(\mathcal{M}-\mathcal{N})v_{|\partial D}=0$ also verifies $(\mathcal{M}v,v)_{L^\partial}\in\mathbb{R}$, we infer using (6a) in (5) that

$$\Re\left((Av,v)_L\right) \geq \mu_0\|v\|_L^2 + \frac{1}{2}(\mathcal{M}v,v)_{L^\partial} \geq \mu_0\|v\|_L^2. \tag{7}$$

Given $f \in L$, our goal is to find a function $u : D \to \mathbb{C}^m$ such that

$$Au = f \ \text{ in } D, \qquad (\mathcal{M} - \mathcal{N})u = 0 \ \text{ on } \partial D. \tag{8}$$

Under the assumptions (1) and (6), Friedrichs proved: (i) the uniqueness of the strong solution $u \in C^1(\overline{D}; \mathbb{C}^m)$ satisfying $(Au, v)_L = (f, v)_L$ for all $v \in L$ and $(\mathcal{M} - \mathcal{N})u = 0$ on ∂D; (ii) the existence of a so-called ultraweak solution $u \in L$ such that $(u, \tilde{A}v)_L = (f, v)_L$ for all $v \in C^1(\overline{D}; \mathbb{C}^m)$ such that $(\mathcal{M}^{\mathsf{H}} + \mathcal{N})v = 0$ on ∂D. In §2, we introduce a mathematical setting relying on boundary operators instead of boundary fields to define a notion of weak solution for (8), and we prove well-posedness of the said formulation by using the BNB Theorem.

1.2. Example 1: advection-reaction equation

Let $\mu \in L^\infty(D; \mathbb{R})$ and let $\boldsymbol{\beta} \in \boldsymbol{L}^\infty(D; \mathbb{R}^d)$ be such that $\nabla \cdot \boldsymbol{\beta} \in L^\infty(D; \mathbb{R})$. Given $f \in L := L^2(D; \mathbb{R})$, we want to find $u : D \to \mathbb{R}$ such that

$$\mu u + \boldsymbol{\beta} \cdot \nabla u = f \quad \text{in } D. \tag{9}$$

This equation models the transport of a solute of concentration u by a flow field with velocity $\boldsymbol{\beta}$, linear reaction coefficient μ ($\mu \geq 0$ corresponds to depletion), and source term f. To recover Friedrichs' formalism, we set $m = 1$, $\mathcal{K} = \mu$, and $\mathcal{A}^k = \beta_k$ for all $k \in \{1{:}d\}$, where $(\beta_k)_{k \in \{1:d\}}$ denote the Cartesian components of $\boldsymbol{\beta}$. The assumption (1a) holds since $\mu \in L^\infty(D; \mathbb{R})$, $\beta_k \in L^\infty(D; \mathbb{R})$ for all $k \in \{1{:}d\}$, and $\mathcal{X} = \nabla \cdot \boldsymbol{\beta} \in L^\infty(D; \mathbb{R})$. The assumption (1b) is trivially satisfied since $m = 1$. Finally, the assumption (1c) is satisfied provided we assume that

$$\mu_0 := \operatorname*{ess\,inf}_{\boldsymbol{x} \in D} (\mu - \tfrac{1}{2} \nabla \cdot \boldsymbol{\beta})(\boldsymbol{x}) > 0. \tag{10}$$

The boundary field is $\mathcal{N} = \boldsymbol{\beta} \cdot \boldsymbol{n}$, and the integration by parts formula (4) is a reformulation of $\int_D ((\nabla \cdot \boldsymbol{\beta}) vw + v(\boldsymbol{\beta} \cdot \nabla w) + w(\boldsymbol{\beta} \cdot \nabla v)) \, \mathrm{d}x = \int_{\partial D} (\boldsymbol{\beta} \cdot \boldsymbol{n}) vw \, \mathrm{d}s$.

To enforce a suitable boundary condition, we need to consider the sign of $(\boldsymbol{\beta} \cdot \boldsymbol{n})$ at the boundary. We define the inflow boundary $\partial D^- = \{\boldsymbol{x} \in \partial D \mid (\boldsymbol{\beta} \cdot \boldsymbol{n})(\boldsymbol{x}) < 0\}$, the outflow boundary $\partial D^+ = \{\boldsymbol{x} \in \partial D \mid (\boldsymbol{\beta} \cdot \boldsymbol{n})(\boldsymbol{x}) > 0\}$, and the characteristic boundary $\partial D^0 = \{\boldsymbol{x} \in \partial D \mid (\boldsymbol{\beta} \cdot \boldsymbol{n})(\boldsymbol{x}) = 0\}$. Then, the inflow boundary condition $u = 0$ on ∂D^- can be enforced by using the boundary field $\mathcal{M} = |\boldsymbol{\beta} \cdot \boldsymbol{n}|$ which satisfies (6). Finally, the L-coercivity property (7) becomes

$$(Av, v)_L \geq \mu_0 \|v\|_L^2 + \frac{1}{2} \int_{\partial D} |\boldsymbol{\beta} \cdot \boldsymbol{n}| v^2 \, \mathrm{d}s.$$

1.3. Example 2: Maxwell's equations

We consider the time-harmonic version of Maxwell's equations in the low-frequency regime where the displacement currents are negligible. Let σ be the electrical conductivity, μ the magnetic permeability, $\omega > 0$ the angular frequency, and $i^2 = -1$. We assume that $\mu, \sigma \in L^\infty(D; \mathbb{R})$, and for simplicity,

that both μ and σ are real-valued. Given $\boldsymbol{j} \in \boldsymbol{L}^2(D) := L^2(D;\mathbb{C}^3)$ and setting $\tilde{\mu} = \omega\mu$, we want to find functions $\boldsymbol{E} : D \to \mathbb{C}^3$ and $\boldsymbol{H} : D \to \mathbb{C}^3$ such that

$$\sigma\boldsymbol{E} - \nabla\times\boldsymbol{H} = \boldsymbol{j} \ \text{ in } D, \qquad i\tilde{\mu}\boldsymbol{H} + \nabla\times\boldsymbol{E} = \boldsymbol{0} \ \text{ in } D. \tag{11}$$

To recover Friedrichs' formalism, we set $m = 6$, $u := (\boldsymbol{E}, \boldsymbol{H})$, $\mathcal{K} = e^{i\theta}\left(\begin{smallmatrix} \sigma\mathbb{I}_3 & \mathbb{O}_3 \\ \mathbb{O}_3 & \tilde{\mu}\mathbb{I}_3 \end{smallmatrix}\right)$ with $\theta = \frac{\pi}{4}$, and $\mathcal{A}^k = \left(\begin{smallmatrix} \mathbb{O}_3 & -e^{i\theta}\mathbb{J}^k \\ e^{-i\theta}\mathbb{J}^k & \mathbb{O}_3 \end{smallmatrix}\right)$, for all $k \in \{1{:}d\}$, where $\mathbb{I}_3$ and $\mathbb{O}_3$ are the identity and null matrix in $\mathbb{C}^{3\times 3}$, respectively, and $\mathbb{J}^k_{ij} = \varepsilon_{ikj}$, for all $i, j, k \in \{1, 2, 3\}$, with ε_{ikj} the Levi-Civita symbol. The assumption (1a) holds since $\sigma, \mu \in L^\infty(D;\mathbb{R})$ and $\mathcal{X}$ is the null matrix in $\mathbb{C}^{6\times 6}$. The assumption (1b) holds since, $\mathbb{J}^k$ being skew-symmetric, we have $(-e^{i\theta}\mathbb{J}^k)^{\mathsf{H}} = -e^{-i\theta}(\mathbb{J}^k)^{\mathsf{T}} = e^{-i\theta}\mathbb{J}^k$. Finally, the assumption (1c) is satisfied provided we assume that

$$\sigma_{\flat,D} := \operatorname*{ess\,inf}_{\boldsymbol{x}\in D} \sigma(\boldsymbol{x}) > 0, \qquad \tilde{\mu}_{\flat,D} := \operatorname*{ess\,inf}_{\boldsymbol{x}\in D} \tilde{\mu}(\boldsymbol{x}) > 0. \tag{12}$$

The boundary field is $\mathcal{N} = \left(\begin{smallmatrix} \mathbb{O}_3 & e^{i\theta}\mathbb{T} \\ -e^{-i\theta}\mathbb{T} & \mathbb{O}_3 \end{smallmatrix}\right)$, where $\mathbb{T}_{ij} = \sum_{k=1}^3 n_k\varepsilon_{ijk}$, for all $i, j \in \{1, 2, 3\}$. Note that the definition of $\mathbb{T}$ implies that $\mathbb{T}\boldsymbol{\xi} = \boldsymbol{\xi}\times\boldsymbol{n}$ for all $\boldsymbol{\xi} \in \mathbb{C}^3$. The integration by parts formula (4) results from $\int_D (\boldsymbol{b}\cdot(\nabla\times\boldsymbol{E}) - \boldsymbol{E}\cdot(\nabla\times\boldsymbol{b}))\,\mathrm{d}x = \int_{\partial D} \boldsymbol{b}\cdot(\boldsymbol{n}\times\boldsymbol{E})\,\mathrm{d}s$.

The boundary conditions $\boldsymbol{H}\times\boldsymbol{n}_{|\partial D} = \boldsymbol{0}$ and $\boldsymbol{E}\times\boldsymbol{n}_{|\partial D} = \boldsymbol{0}$ can be enforced, respectively, by using the boundary fields $\mathcal{M}_H = \left(\begin{smallmatrix} \mathbb{O}_3 & -e^{i\theta}\mathbb{T} \\ -e^{-i\theta}\mathbb{T} & \mathbb{O}_3 \end{smallmatrix}\right)$ and $\mathcal{M}_E = \left(\begin{smallmatrix} \mathbb{O}_3 & e^{i\theta}\mathbb{T} \\ e^{-i\theta}\mathbb{T} & \mathbb{O}_3 \end{smallmatrix}\right)$, which both satisfy (6), and the coercivity property (7) becomes

$$\Re(A(\boldsymbol{E}, \boldsymbol{H}), (\boldsymbol{E}, \boldsymbol{H}))_{L^2(D;\mathbb{C}^6)} \geq \frac{1}{\sqrt{2}}\left(\sigma_{\flat,D}\|\boldsymbol{E}\|^2_{\boldsymbol{L}^2(D)} + \tilde{\mu}_{\flat,D}\|\boldsymbol{H}\|^2_{\boldsymbol{L}^2(D)}\right).$$

2. Weak formulation and well-posedness for Friedrichs' systems

The aim of this section is to devise a weak formulation of Friedrichs' systems for which well-posedness can be established by using the Banach–Nečas–Babuška (BNB) Theorem which provides necessary and sufficient conditions for well-posedness in the form of inf-sup conditions, see [12, Thm. 2.6]. The material is inspired from a series of papers by the authors [13, 14].

2.1. The graph space

We consider the space $S := C_0^\infty(D;\mathbb{C}^m)$, composed of the smooth $\mathbb{C}^m$-valued fields compactly supported in D, and the Hilbert space $L := L^2(D;\mathbb{C}^m)$, which we use as pivot space (i.e., $L \equiv L'$). While other functional settings could be considered, we will see in the forthcoming sections that L^2 plays a prominent role in a large class of stabilized finite element techniques.

The operators A and $\tilde{A}$ defined in (2) and (3), respectively, are each bounded in S with values in L and the following holds: There is c such that

$$(A\phi, \psi)_L = (\phi, \tilde{A}\psi)_L, \quad \forall\phi, \psi \in S, \tag{13a}$$

$$\|(A + \tilde{A})\phi\|_L \leq c\,\|\phi\|_L \quad \forall\phi \in S. \tag{13b}$$

The equality (13a) follows from Lemma 1.1, while (13b) follows from the definitions of A and $\tilde{A}$ and the boundedness property (1a). Let us define the inner product $(\cdot,\cdot)_V := \mu_0(\cdot,\cdot)_L + \mu_0^{-1}(A_1(\cdot), A_1(\cdot))_L$ and let the induced norm be denoted by $\|\cdot\|_V$ with $\|\cdot\|_V^2 = \mu_0\|\cdot\|_L^2 + \mu_0^{-1}\|A_1(\cdot)\|_L^2$ (the scaling factors μ_0 and μ_0^{-1} are introduced so that both terms have coherent units).

Let V_S be the completion of S with respect to the norm $\|\cdot\|_V$, i.e., $V_S = \overline{S}^V$. Using L as pivot space leads to $S \subset V_S \hookrightarrow L \equiv L' \hookrightarrow V_S' \subset S'$, where S' is the algebraic dual of S and L', V_S' are topological duals. By density, the operators A and $\tilde{A}$ can be extended to bounded linear operators from V_S to L; we say that V_S is the *minimal domain* of A and $\tilde{A}$. Owing to (13), we infer by density that $(A\phi,\psi)_L = (\phi,\tilde{A}\psi)_L$, for all $\phi,\psi \in V_S$. Let now $v \in L$; then, Av can be defined in V_S' by setting $\langle Av,\phi\rangle_{V_S',V_S} = (v,\tilde{A}\phi)_L$, for all $\phi \in V_S$. This definition allows us to extend A to a bounded linear operator from L to V_S'. Similarly we define $\langle \tilde{A}v,\phi\rangle_{V_S',V_S} = (v,A\phi)_L$, for all $v \in L$ and all $\phi \in V_S$. Since $L \subset V_S'$, it makes sense to define the *graph space* (or *maximal domain* of A and $\tilde{A}$) as

$$V := \{v \in L;\ A_1 v \in L\}. \tag{14}$$

By construction, $A \in \mathcal{L}(V;L)$, $\tilde{A} \in \mathcal{L}(V;L)$.

Proposition 2.1 (Hilbert space)**.** *The graph space V is a Hilbert space when equipped with the inner product $(\cdot,\cdot)_V$. The norm $\|\cdot\|_V$ is called the graph norm.*

2.2. The boundary operators

Since A_1 is a first-order differential operator, defining the trace at the boundary of a function in the graph space V is not straightforward. The trace can be given a meaning in $H^{-\frac{1}{2}}(\partial D;\mathbb{C}^m)$, see Rauch [27]. However, this meaning is not suitable for the weak formulation we have in mind; this is why we now introduce two additional operators N and M to replace the boundary fields $\mathcal{N}$ and $\mathcal{M}$. We define the operator $N \in \mathcal{L}(V;V')$ by (compare with (4))

$$\langle Nv,w\rangle_{V',V} := (Av,w)_L - (v,\tilde{A}w)_L, \qquad \forall v,w \in V. \tag{15}$$

This definition makes sense since both A and $\tilde{A}$ are in $\mathcal{L}(V;L)$. Moreover, the operator N is self-adjoint since (15) can be rewritten as

$$\langle Nv,w\rangle_{V',V} = (\mathcal{X}v,w)_L + (A_1 v,w)_L + (v,A_1 w)_L, \tag{16}$$

so that $\langle Nv,w\rangle_{V',V} = \overline{\langle Nw,v\rangle_{V',V}}$. Furthermore, we have $V_S \subset \ker(N)$ and $\mathrm{im}(N) \subset V_S^\perp = \{v' \in V' \mid \forall \phi \in V_S, \langle v',\phi\rangle_{V_S',V_S} = 0\}$. Actually, as proved in [17], the following holds: $\ker(N) = V_S$, $\mathrm{im}(N) = V_S^\perp$. The fact that $\ker(N) = V_S$ means that N is a *boundary operator*.

Boundary conditions in Friedrichs' systems can be formulated by assuming that there exists an operator $M \in \mathcal{L}(V;V')$ such that

$$M \text{ is monotone, i.e., } \Re(\langle Mv,v\rangle_{V',V}) \geq 0 \text{ for all } v \in V, \tag{17a}$$

$$\ker(N-M) + \ker(N+M) = V. \tag{17b}$$

Let $M^* \in \mathcal{L}(V;V')$ denote the adjoint operator of M, so that $\langle M^*w, v\rangle_{V',V} = \overline{\langle Mv, w\rangle_{V',V}}$. It is proved in [17] that, under the assumptions (17),

$$\ker(N) = \ker(M) = \ker(M^*), \quad \text{and} \quad \mathrm{im}(N) = \mathrm{im}(M) = \mathrm{im}(M^*).$$

In particular, M is a *boundary operator*, just like N.

2.3. Well-posedness

Given $f \in L$, the problem we want to solve (compare with (8)) is to find

$$u \in V_0 := \ker(M-N) \text{ such that } Au = f \text{ in } L. \tag{18}$$

To recast this problem into a weak form, we introduce the sesquilinear form $a(v,w) := (Av, w)_L$, for all $(v,w) \in V \times L$. Letting $\ell(w) := (f,w)_L$, we consider the following weak problem:

$$\begin{cases} \text{Find } u \in V_0 \text{ such that} \\ a(u,w) = \ell(w), \quad \forall w \in L. \end{cases} \tag{19}$$

Theorem 2.2 (Well-posedness)**.** *Let N be defined by* (15) *and assume* (1)*, then $\Re(a(v,v)) \geq \mu_0\|v\|_L^2 + \frac{1}{2}\langle Nv, v\rangle_{V',V}, \forall v \in V$. Moreover, let M satisfy* (17)*, then $\Re(a(v,v)) \geq \mu_0\|v\|_L^2 + \frac{1}{2}\Re(\langle Mv, v\rangle_{V',V}) \geq \mu_0\|v\|_L^2, \forall v \in V_0$. If* (1) *and* (17) *hold, then the model problem* (19) *is well-posed, i.e., $A : V_0 \to L$ is an isomorphism.*

Remark 2.3 (Positivity assumption (1c))**.** This assumption can be relaxed if the missing control on $\|v\|_L$ can be recovered from an estimate on $\|A_1 v\|_L$. This is possible in the context of elliptic PDEs in mixed form by invoking a Poincaré-type inequality. Furthermore, everything that is said hereafter holds true by assuming that $A = K + A_1$ where K is a bounded operator on L satisfying the assumption $((K+K^*)v - \mathcal{X}v, v)_L \geq 2\mu_0\|v\|_L^2$. The formal adjoint is then defined by $\tilde{A}v = K^*v - \mathcal{X}v + A_1 v$. For instance let $D = (0,a)\times(-1,1)$, $a > 0$, and let $K : L \to L$, with $L = L^2(D;\mathbb{R})$, be such that $Kv(x,y) = v(x,y) - \frac{\sigma}{2}\int_{-1}^{+1} v(x,\xi)\,\mathrm{d}\xi$ where $\sigma \in [0,1)$. Then $((K+K^*)v, v)_L = 2(Kv,v)_L \geq 2\|v\|_L^2 - 2\sigma\|v\|_L^2 = 2(1-\sigma)\|v\|_L^2$. This is the type of structure one encounters when solving the neutron transport equation.

Example 2.4 (Advection-reaction)**.** The bilinear form a is

$$a(v,w) = \int_D (\mu vw + (\boldsymbol{\beta}\cdot\nabla v)w)\,\mathrm{d}x, \qquad \forall v \in V,\ \forall w \in L^2(D;\mathbb{R}),$$

with $V = \{v \in L^2(D;\mathbb{R}) \mid \boldsymbol{\beta}\cdot\nabla v \in L^2(D;\mathbb{R})\}$. Moreover,

$$\langle Nv, w\rangle_{V',V} = \int_D ((\nabla\cdot\boldsymbol{\beta})vw + w(\boldsymbol{\beta}\cdot\nabla v) + v(\boldsymbol{\beta}\cdot\nabla w))\,\mathrm{d}x.$$

A result on traces of functions in V is needed to link N with $\mathcal{N} = \boldsymbol{\beta}\cdot\boldsymbol{n}$. Such a result is not straightforward, since the trace theorem for functions in $H^s(D;\mathbb{R})$,

$s > \frac{1}{2}$, cannot be applied. It is shown in [14] that if the inflow and outflow boundaries are well-separated, i.e., $\min_{(\boldsymbol{x},\boldsymbol{y})\in\partial D^-\times\partial D^+} \|\boldsymbol{x}-\boldsymbol{y}\|_{\ell^2(\mathbb{R}^d)} > 0$, then the trace operator $\gamma : C^0(\overline{D}) \to C^0(\partial D)$ such that $\gamma(v) = v_{|\partial D}$ can be extended to a bounded linear operator from V to $L^2_{|\boldsymbol{\beta}\cdot\boldsymbol{n}|}(\partial D;\mathbb{R})$, where the subscript $|\boldsymbol{\beta}\cdot\boldsymbol{n}|$ means that the measure $\mathrm{d}s$ is replaced by $|\boldsymbol{\beta}\cdot\boldsymbol{n}|\,\mathrm{d}s$. This result implies that $\langle Nv, w\rangle_{V',V} = \int_{\partial D} \mathcal{N}vw\,\mathrm{d}s$ for all $v, w \in V$. Furthermore, the inflow boundary condition $u = 0$ on ∂D^- can be enforced by means of the boundary operator $M \in \mathcal{L}(V;V')$ defined by $\langle Mv, w\rangle_{V',V} = \int_{\partial D} |\boldsymbol{\beta}\cdot\boldsymbol{n}|vw\,\mathrm{d}s$, which satisfies (17). Note that the separation assumption cannot be circumvented if one wishes to work with traces in $L^2_{|\boldsymbol{\beta}\cdot\boldsymbol{n}|}(\partial D;\mathbb{R})$, regardless of the regularity of $\boldsymbol{\beta}$. For instance, let $D = \{(x_1,x_2) \in \mathbb{R}^2 \mid 0 < x_2 < 1 \text{ and } |x_1| < x_2\}$ with $\boldsymbol{\beta} = (1,0)^\mathsf{T}$. One can verify that the function $u(x_1,x_2) = x_2^\alpha$ is in V for $\alpha > -1$, but $u_{|\partial D} \in L^2(|\boldsymbol{\beta}\cdot\boldsymbol{n}|;\partial D)$ only if $\alpha > -\frac{1}{2}$. □

Example 2.5 (Maxwell)**.** The sesquilinear form a is

$$a(v,w) = \int_D \left(e^{i\theta}\sigma\boldsymbol{E}\cdot\overline{\boldsymbol{e}} + ie^{-i\theta}\tilde{\mu}\boldsymbol{H}\cdot\overline{\boldsymbol{b}} - e^{i\theta}(\nabla\times\boldsymbol{H})\cdot\overline{\boldsymbol{e}} + e^{-i\theta}(\nabla\times\boldsymbol{E})\cdot\overline{\boldsymbol{b}}\right)\mathrm{d}x,$$

for all $v = (\boldsymbol{E},\boldsymbol{H}) \in V$ and all $w = (\boldsymbol{e},\boldsymbol{b}) \in L$ (note that we use the Euclidean dot product and write the complex conjugate explicitly), with $V = \boldsymbol{H}(\mathrm{curl};D)\times\boldsymbol{H}(\mathrm{curl};D)$, $\boldsymbol{H}(\mathrm{curl};D) = \{\boldsymbol{A} \in L^2(D;\mathbb{C}^3); \nabla\times\boldsymbol{A} \in L^2(D;\mathbb{C}^3)\}$, and $L = L^2(D;\mathbb{C}^6)$. Moreover,

$$\langle N(\boldsymbol{E},\boldsymbol{H}),(\boldsymbol{e},\boldsymbol{b})\rangle_{V',V} = e^{i\theta}t(\boldsymbol{H},\boldsymbol{e}) - e^{-i\theta}t(\boldsymbol{E},\boldsymbol{h}),$$

where $t(\boldsymbol{A},\boldsymbol{a}) = \int_D(\boldsymbol{A}\cdot(\nabla\times\overline{\boldsymbol{a}}) - (\nabla\times\boldsymbol{A})\cdot\overline{\boldsymbol{a}})\,\mathrm{d}x$. Since $\boldsymbol{E}\times\boldsymbol{n}$ and $\boldsymbol{H}\times\boldsymbol{n}$ are in $\boldsymbol{H}^{-1/2}(\partial D)$, if $\boldsymbol{e}$ and $\boldsymbol{b}$ are in $\boldsymbol{H}^1(D)$, we have $\langle N(\boldsymbol{E},\boldsymbol{H}),(\boldsymbol{e},\boldsymbol{b})\rangle_{V',V} = e^{i\theta}\langle\boldsymbol{H}\times\boldsymbol{n},\boldsymbol{e}\rangle_{\boldsymbol{H}^{-\frac{1}{2}},\boldsymbol{H}^{\frac{1}{2}}} - e^{-i\theta}\langle\boldsymbol{E}\times\boldsymbol{n},\boldsymbol{b}\rangle_{\boldsymbol{H}^{-\frac{1}{2}},\boldsymbol{H}^{\frac{1}{2}}}$. The boundary condition $\boldsymbol{H}\times\boldsymbol{n} = \boldsymbol{0}$ can be enforced by means of the boundary operator $\langle M(\boldsymbol{E},\boldsymbol{H}),(\boldsymbol{e},\boldsymbol{b})\rangle_{V',V} = -e^{i\theta}t(\boldsymbol{H},\boldsymbol{e}) - e^{-i\theta}t(\boldsymbol{E},\boldsymbol{h})$, which satisfies (17). □

3. Residual-based stabilization

This section is concerned with the approximation of Friedrichs' systems using H^1-conforming finite elements in a standard Galerkin setting. The main issue one faces in this context is to achieve stability. At the continuous level, the proof of Theorem 2.2 shows that one needs to consider the first-order derivative A_1v as test function to control the graph norm of a function v. Unfortunately, this possibility is lost when working with H^1-conforming finite elements since the first-order derivative of v can no longer be represented by discrete test functions. As a result, one needs to devise suitable stabilization mechanisms. Those presented in this section are inspired by the Least-Squares (LS), or minimial residual, technique from linear algebra. The LS approximation gives optimal error estimates in the graph norm, but, unfortunately, gives suboptimal L^2-error

estimates in most situations. The Galerkin/Least-Squares (GaLS) method improves the situation by combining the standard Galerkin approach with the LS technique and mesh-dependent weights. GaLS gives quasi-optimal L^2-error estimates and optimal graph-norm estimates. We further improve GaLS in the next section by introducing a boundary penalty technique that enforces boundary conditions weakly in the spirit of the theory of Friedrichs' sytstems.

3.1. Least-Squares formulation

Given $f \in L$, let us consider the model problem (19). This problem is well-posed, see Theorem 2.2. The LS version of problem (19) is the following:

$$\left\{ \begin{array}{l} \text{Find } u \in V_0 \text{ such that} \\ a^{\text{LS}}(u,w) := (Au, Aw)_L = (f, Aw)_L, \quad \forall w \in V_0. \end{array} \right. \tag{20}$$

Observe that the test space is the same as the solution space in (20). Since $A : V_0 \to L$ is an isomorphism, requiring that $(Au, Aw)_L = (f, Aw)_L$ for all $w \in V_0$ is equivalent to ask that $(Au, w)_L = (f, w)_L$ for all $w \in L$. Hence, the problems (19) and (20) are equivalent. Actually, the well-posedness of (20) is a direct consequence of the Lax–Milgram Lemma, since there are real numbers $0 < \alpha \leq \varpi < \infty$ such that $\alpha \|v\|_V \leq \|Av\|_L \leq \varpi \|v\|_V$ for all $v \in V_0$.

Proposition 3.1 (V_0-coercivity)**.** *a^{LS} is bounded and coercive on V_0.*

Remark 3.2 (Minimal residual)**.** Consider the functional $\mathfrak{J} : V_0 \to \mathbb{R}$ defined by $\mathfrak{J}(v) := \frac{1}{2}\|Av - f\|_L^2$ for all $v \in V_0$. The Fréchet derivative of $\mathfrak{J}$ is such that $D\mathfrak{J}(v)(w) = \Re((Av - f, Aw)_L)$ for all $w \in V_0$, i.e., the problem (20) amounts to $D\mathfrak{J}(v) = 0$. Since the functional $\mathfrak{J}$ is strictly convex, the solution u of (20) is the global minimizer of $\mathfrak{J}$ over V_0. This LS technique is well-known in the linear algebra context where it can be traced back to Gauss and Legendre. Starting from the linear system $\mathcal{A}U = B$ with $\mathcal{A}$ invertible and multiplying by $\mathcal{A}^{\mathsf{H}}$ leads to the so-called normal equations $(\mathcal{A}^{\mathsf{H}}\mathcal{A})U = \mathcal{A}^{\mathsf{H}}B$ where the matrix $\mathcal{A}^{\mathsf{H}}\mathcal{A}$ is Hermitian positive-definite. □

3.2. Least-Squares approximation using Finite Elements

We assume that, for all $h > 0$, we have at hand a finite-dimensional space $V_{h0} \subset V_0$ built by using a shape-regular mesh sequence $(\mathcal{T}_h)_{h>0}$ and a finite element of degree $k \geq 1$. For simplicity, we consider the equal-order case for all the solution components. The space V_{h0} is H^1-conforming and composed of continuous, piecewise polynomial functions in $\overline{D}$. Let us assume now that we have at hand a quasi-interpolation operator $\mathcal{I}_{h0} : V_0 \to V_{h0}$ with optimal local approximation properties: There is a uniform constant c such that

$$\|v - \mathcal{I}_{h0}(v)\|_{L(K)} + h_K \|\nabla(v - \mathcal{I}_{h0}(v))\|_{\boldsymbol{L}(K)} \leq c\, h_K^{1+r} |v|_{H^{1+r}(D_K, \mathbb{C}^m)}, \tag{21}$$

for all $r \in [0, k]$, all $v \in H^{1+r}(D, \mathbb{C}^m) \cap V_0$, and all $K \in \mathcal{T}_h$, with $L(K) := L^2(K; \mathbb{C}^m)$ and where D_K is the interior of the set composed of all the mesh cells having a non-empty intersection with K.

We construct a discrete counterpart of (20) as follows:

$$\begin{cases} \text{Find } u_h \in V_{h0} \text{ such that} \\ a^{\mathrm{LS}}(u_h, w_h) = (f, Aw_h)_L, \quad \forall w_h \in V_{h0}. \end{cases} \tag{22}$$

Theorem 3.3 (Well-posedness and error bound)**.** *The problem* (22) *has a unique solution* u_h*, and the following error bound holds:*

$$\|u - u_h\|_V \le \frac{\varpi}{\alpha} \inf_{v_h \in V_{h0}} \|u - v_h\|_V. \tag{23}$$

Using (21), we infer the following approximation result in the graph norm: $\|u - \mathcal{I}_{h0}(u)\|_V \le c\,\mu_0^{-\frac{1}{2}} \phi_D h^r |u|_{H^{1+r}(D;\mathbb{C}^m)}$, with $\phi_D := \max(\beta_D, \mu_0 h)$ and $\beta_D = \max_{k\in\{1:d\}} \|\mathcal{A}^k\|_{L^\infty(D;\mathbb{C}^{m\times m})}$. Assuming $u \in H^{1+r}(D;\mathbb{C}^m)$ and using the above approximation result, we infer that

$$\mu_0^{\frac{1}{2}} \|u - u_h\|_L + \mu_0^{-\frac{1}{2}} \|A_1(u - u_h)\|_L \le c\,\mu_0^{-\frac{1}{2}} \phi_D h^r |u|_{H^{1+r}(D;\mathbb{C}^m)}. \tag{24}$$

When $r = k$, the estimate on $\|A_1(u - u_h)\|_L$ is optimal, but the estimate on $\|u - u_h\|_L$ is *suboptimal* by one order. It is sometimes possible to improve the L-norm error estimate by means of the Aubin–Nitsche duality argument, but this is not systematic since, very often, first-order PDEs do not have a smoothing property. For instance, this improvement is possible for the one-dimensional transport equation and for Darcy's equation.

The LS technique has gained popularity in the numerical analysis community at the beginning of the 1970s following a series of papers by Bramble and Schatz [3, 4], although it was already popular in the Russian literature (see Džiškariani [11], Lučka [24]).

3.3. Galerkin/Least-Squares

In this section, we devise and analyze a Galerkin Least-Squares (GaLS) approximation of the model problem (19) introduced in Hughes et al. [22]. A nonsymmetric variant known under the names Streamline Upwind Petrov–Galerkin (SUPG) or streamline diffusion method has been introduced in Brooks and Hughes [5] and analyzed in Johnson et al. [23], see Example 3.6.

We define the following local quantities:

$$\beta_K = \max_{k\in\{1:d\}} \|\mathcal{A}^k\|_{L^\infty(K;\mathbb{C}^{m\times m})}, \tag{25}$$

$$\tau_K = \big(\max(\beta_K h_K^{-1}, \mu_0)\big)^{-1} = \min(\beta_K^{-1} h_K, \mu_0^{-1}), \tag{26}$$

for all $K \in \mathcal{T}_h$, where μ_0 is defined in (1c) (the second equality is meaningful if β_K is nonzero; if $\beta_K = 0$, then $\tau_K = \mu_0^{-1}$). For instance, for the advection-reaction equation, μ_0 is the reciprocal of a time, β_K is a local velocity, and τ_K is a local time scale. With a slight abuse of notation, we define the piecewise constant function $\tau : D \to \mathbb{R}$ such that $\tau_{|K} = \tau_K$ for all $K \in \mathcal{T}_h$. In what follows, we consider the Euclidean (or Hermitian) norm denoted $\|\cdot\|_{\ell^2}$ for $\mathbb{C}^{m\times m}$-valued

fields, we set $\|\cdot\|_{L^\infty(D;\mathbb{C}^{m\times m})} = \|\,\|\cdot\|_{\ell^2}\,\|_{L^\infty(D;\mathbb{R})}$, and we assume for simplicity that

$$\max(\|\mathcal{K}\|_{L^\infty(D;\mathbb{C}^{m\times m})}, \|\mathcal{X}\|_{L^\infty(D;\mathbb{C}^{m\times m})}) \le c_{\mathcal{K},\mathcal{X}}\mu_0, \tag{27}$$

and we hide the factor $c_{\mathcal{K},\mathcal{X}}$ in the generic constants used in the error analysis.

We consider the finite element setting of §3.2. We define the following discrete sesquilinear forms on $V_{h0}\times V_{h0}$:

$$a_h(v_h,w_h) := (Av_h,w_h)_L + r_h(v_h,w_h), \qquad r_h(v_h,w_h) := (Av_h,\tau Aw_h)_L. \tag{28}$$

The sesquilinear form $(Av_h,w_h)_L$ is the Galerkin part of the formulation and the term $r_h(v_h,w_h)$ is the least-squares part. The role of r_h is to stabilize the formulation. We consider the following discrete problem:

$$\begin{cases} \text{Find } u_h\in V_{h0} \text{ such that} \\ a_h(u_h,w_h) = \ell_h(w_h) := (f, w_h+\tau Aw_h)_L, \quad \forall w_h\in V_{h0}. \end{cases} \tag{29}$$

As usual the four steps of the analysis consist of (i) establishing stability, (ii) estimating the consistency error, (iii) proving a boundedness estimate, and (iv) using the approximation properties of finite elements. We set $V_\flat = V_0 + V_{h0}$ and observe that $V_\flat = V_0$ since the approximation is V_0-conforming. Proceeding in the spirit of Strang's Second Lemma for the error analysis, we extend the sesquilinear form a_h to $V_0\times V_{h0}$, and we equip the space V_0 with the norms:

$$\|v\|_{V_\flat}^2 := \mu_0\|v\|_L^2 + \frac{1}{2}|v|_M^2 + \|\tau^{\frac12}Av\|_L^2, \qquad \|v\|_{V_{\flat\sharp}}^2 := \|v\|_{V_\flat}^2 + \|\tau^{-\frac12}v\|_L^2, \tag{30}$$

with the boundary semi-norm $|v|_M^2 := \Re(\langle Mv,v\rangle_{V',V})$.

Theorem 3.4 (Convergence)**.** *(i) The discrete sesquilinear form a_h satisfies $\Re(a_h(v_h,v_h)) \ge \|v_h\|_{V_\flat}^2$, for all $v_h\in V_{h0}$. Consequently, the discrete problem* (29) *is well-posed. (ii) The discrete problem* (29) *is exactly consistent. (iii) There is c, uniform with respect to h, such that, $|a_h(v,w_h)| \le c\,\|v\|_{V_{\flat\sharp}}\|w_h\|_{V_\flat}$ for all $(v,w_h)\in V_0\times V_{h0}$. (iv) Let u be the unique solution to* (19) *and let u_h be the unique solution to* (29)*. There is c, uniform with respect to h, such that*

$$\|u-u_h\|_{V_\flat} \le c \inf_{v_h\in V_{h0}} \|u-v_h\|_{V_{\flat\sharp}}. \tag{31}$$

Moreover $\|u-u_h\|_{V_\flat}^2 \le c\sum_{K\in\mathcal{T}_h}\max(\beta_K,\mu_0 h_K)h_K^{2r+1}|u|^2_{H^{1+r}(D_K;\mathbb{C}^m)}$ if $u\in H^{1+r}(D;\mathbb{C}^m)$, for all $r\in[0,k]$. Upon denoting $\phi_D := \max(\beta_D,\mu_0 h)$ and $\beta_D := \max_{K\in\mathcal{T}_h}\beta_K$, this implies in particular that $\|u-u_h\|_{V_\flat} \le c\,\phi_D^{\frac12}h^{r+\frac12}|u|_{H^{1+r}(D;\mathbb{C}^m)}$.

Assuming $u\in H^{k+1}(D;\mathbb{C}^m)$, the above result implies that

$$\mu_0^{\frac12}\|u-u_h\|_L + \|\tau^{\frac12}A_1(u-u_h)\|_L \le c\,\phi_D^{\frac12}h^{k+\frac12}|u|_{H^{k+1}(D;\mathbb{C}^m)}.$$

Observe that the estimate on $\|u-u_h\|_L$ is improved by half a power in h when compared to that obtained with the LS technique, and the estimate on $\|A_1(u-u_h)\|_L$ is now a localized version of the LS estimate (24).

Example 3.5 (Advection-reaction)**.** Consider the PDE $\mu u + \boldsymbol{\beta}\cdot\nabla u = f$ with the inflow boundary condition $u = 0$ on ∂D^-, see §1.2. Assume that all the mesh boundary faces are a subset of either ∂D^- or $\partial D \setminus \partial D^-$. Let $P_k^{\rm g}(\mathcal{T}_h)$ be the H^1-conforming finite element space constructed on the mesh $\mathcal{T}_h$ using finite elements of degree $k \geq 1$ [16]. Set $V_{h0} := \{v_h \in P_k^{\rm g}(\mathcal{T}_h) \mid v_{h|\partial D^-} = 0\}$. The GaLS discretization consists of seeking $u_h \in V_{h0}$ such that

$$\int_D (\mu u_h + \boldsymbol{\beta}\cdot\nabla u_h) w_h \,\mathrm{d}x + \int_D \tau(\mu u_h + \boldsymbol{\beta}\cdot\nabla u_h)(\mu w_h + \boldsymbol{\beta}\cdot\nabla w_h)\,\mathrm{d}x = \ell_h(w_h),$$

for all $w_h \in V_{h0}$, with $\tau_K = \min(\beta_K^{-1} h_K, \mu_0^{-1})$, $\beta_K = \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(K)}$, and with right-hand side $\ell_h(w_h) = \int_D f w_h \,\mathrm{d}x + \int_D \tau f(\mu w_h + \boldsymbol{\beta}\cdot\nabla w_h)\,\mathrm{d}x$. Provided $u \in H^{1+r}(D)$, $r \in [0,k]$, and with $\phi_D := \max(\|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}, \mu_0 h)$, Theorem 3.4 gives

$$\mu_0^{\frac12}\|u - u_h\|_{L^2(D)} + \|\tau^{\frac12}\boldsymbol{\beta}\cdot\nabla(u - u_h)\|_{L^2(D)} \leq c\,\phi_D^{\frac12} h^{r+\frac12}|u|_{H^{1+r}(D)}. \qquad \square$$

Example 3.6 (SUPG)**.** Assume that $h_K \leq \beta_K \mu_0^{-1}\min(1, \frac12 \frac{\mu_0^2}{\mu_\infty^2})$ with $\mu_\infty = \|\mathcal{K}\|_{L^\infty(D;\mathbb{C}^{m\times m})}$, for all $K \in \mathcal{T}_h$. The same error estimate as in the GaLS approximation is obtained by considering the following discrete problem: Find $u_h \in V_{h0}$ such that $a_h^{\rm SUPG}(u_h, w_h) = (f, w_h + \tau A_1 w_h)_L$ for all $w_h \in V_{h0}$ with the SUPG-stabilized sesquilinear form $a_h^{\rm SUPG}(v_h, w_h) = (Av_h, w_h)_L + (Av_h, \tau A_1 w_h)_L$. $\square$

Example 3.7 (Maxwell)**.** Consider the PDEs $\sigma\boldsymbol{E} - \nabla\times\boldsymbol{H} = \boldsymbol{f}$ and $i\tilde\mu\boldsymbol{H} + \nabla\times\boldsymbol{E} = \boldsymbol{0}$ with the boundary condition $\boldsymbol{H}\times\boldsymbol{n} = \boldsymbol{0}$. Define the reference length scale $\ell_* = (\sigma_{\flat,D}\tilde\mu_{\flat,D})^{-\frac12}$. Set $\boldsymbol{W}_h := \boldsymbol{P}_k^{\rm g}(\mathcal{T}_h)$ and $\boldsymbol{W}_{h0} := \{\boldsymbol{b}_h \in \boldsymbol{W}_h \mid \boldsymbol{b}_h\times\boldsymbol{n}_{|\partial D} = \boldsymbol{0}\}$. The GaLS approximation amounts to finding $(\boldsymbol{E}_h, \boldsymbol{H}_h) \in V_{h0} := \boldsymbol{W}_h\times\boldsymbol{W}_{h0}$ such that

$$\begin{aligned}
&\int_D \big((\sigma\boldsymbol{E}_h - \nabla\times\boldsymbol{H}_h)\cdot\overline{\boldsymbol{e}}_h + (i\tilde\mu\boldsymbol{H}_h + \nabla\times\boldsymbol{E}_h)\cdot\overline{\boldsymbol{b}}_h\big)\,\mathrm{d}x\\
&+ \int_D \tilde\mu_{\flat,D}^{-1}\tau(i\tilde\mu\boldsymbol{H}_h + \nabla\times\boldsymbol{E}_h)\cdot(-i\tilde\mu\overline{\boldsymbol{b}}_h + \nabla\times\overline{\boldsymbol{e}}_h)\,\mathrm{d}x\\
&+ \int_D \sigma_{\flat,D}^{-1}\tau(\sigma\boldsymbol{E}_h - \nabla\times\boldsymbol{H}_h)\cdot(\sigma\overline{\boldsymbol{e}}_h - \nabla\times\overline{\boldsymbol{b}}_h)\,\mathrm{d}x = \ell_h(w_h),
\end{aligned}$$

for all $w_h = (\boldsymbol{e}_h, \boldsymbol{b}_h) \in V_{h0}$, with local weights $\tau_K = \min(\ell_*^{-1} h_K, 1)$, and right-hand side $\ell_h(w_h) = \int_D \boldsymbol{j}\cdot\overline{\boldsymbol{e}}_h \,\mathrm{d}x + \int_D \sigma_{\flat,D}^{-1}\tau\boldsymbol{j}\cdot(\sigma\overline{\boldsymbol{e}}_h - \nabla\times\overline{\boldsymbol{b}}_h)\,\mathrm{d}x$. Provided $(\boldsymbol{E}, \boldsymbol{H}) \in \boldsymbol{H}^{1+r}(D)\times\boldsymbol{H}^{1+r}(D)$, $r \in [0,k]$, Theorem 3.3, combined with the approximation properties of V_{h0}, yields

$$\begin{aligned}
&\sigma_{\flat,D}^{\frac12}\|\boldsymbol{E} - \boldsymbol{E}_h\|_{\boldsymbol{L}^2(D)} + \tilde\mu_{\flat,D}^{\frac12}\|\boldsymbol{H} - \boldsymbol{H}_h\|_{\boldsymbol{L}^2(D)} + \tilde\mu_{\flat,D}^{-\frac12}\|\tau^{\frac12}\nabla\times(\boldsymbol{E} - \boldsymbol{E}_h)\|_{\boldsymbol{L}^2(D)}\\
&\quad + \sigma_{\flat,D}^{-\frac12}\|\tau^{\frac12}\nabla\times(\boldsymbol{H} - \boldsymbol{H}_h)\|_{\boldsymbol{L}^2(D)} \leq c\phi_D^{\frac12} h^{r+\frac12}\big(|\boldsymbol{E}|_{\boldsymbol{H}^{1+r}(D)} + |\boldsymbol{H}|_{\boldsymbol{H}^{1+r}(D)}\big),
\end{aligned}$$

with $\phi_D = \max(\ell_*, \mu_0 h)$. $\square$

4. Boundary penalty for Friedrichs' systems

It is not always possible, or easy, to build V_0-conforming finite elements; think for instance of a boundary condition enforcing the value of the normal or tangential component of a vector field at the boundary of a domain that is not a rectangular parallelepiped. The goal of this section is twofold: First, to show how to enforce boundary conditions weakly in Friedrichs' systems; second, to combine this approach with the GaLS stabilization. The boundary penalty technique introduced herein will be used again in Section 5.

4.1. Model problem

We now consider the sesquilinear form

$$\widetilde{a}(v,w) := (Av,w)_L + \frac{1}{2}\langle (M-N)v, w\rangle_{V',V}, \qquad \forall v,w \in V. \tag{32}$$

The last term on the right-hand side is used to enforce the boundary condition $u \in \ker(M-N)$ weakly. Owing to this additional term, the test functions are now restricted to be in the graph space V; i.e., taking test functions in L is no longer legitimate. The model problem that we consider is the following:

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ \widetilde{a}(u,w) = (f,w)_L, \quad \forall w \in V. \end{cases} \tag{33}$$

If u solves (33), taking w in $C_0^\infty(D;\mathbb{C}^m)$ implies that $Au = f$ in $L^2(D;\mathbb{C}^m)$; then, we have $\langle (M-N)u, w\rangle_{V',V} = 0$ for all $w \in V$, i.e., $u \in \ker(M-N)$.

Lemma 4.1 (L-coercivity and well-posedness)**.** *The sesquilinear form $\widetilde{a}$ defined by* (32) *is such that $\Re(\widetilde{a}(v,v)) \geq \mu_0\|v\|_L^2 + \frac{1}{2}|v|_M^2$, for all $v \in V$. Problem* (33) *is well-posed, and its unique solution is the unique solution to* (19)*.*

4.2. Boundary penalty method

We are interested in a V-conforming approximation of the model problem (33). For this purpose, we assume that, for all $h > 0$, we have at hand an H^1-conforming finite-dimensional space $V_h \subset V$, built by using a shape-regular mesh sequence $(\mathcal{T}_h)_{h>0}$ and a finite element of degree $k \geq 1$, and a quasi-interpolation operator $\mathcal{I}_h : V \to V_h$ with optimal local approximation properties: There is a uniform constant c such that

$$\|v - \mathcal{I}_h(v)\|_{L(K)} + h_K\|\nabla(v - \mathcal{I}_h(v))\|_{\boldsymbol{L}(K)} \leq c\, h_K^{1+r}|v|_{H^{1+r}(D_K,\mathbb{C}^m)}, \tag{34}$$

for all $r \in [0,k]$, all $v \in H^{1+r}(D,\mathbb{C}^m)$, and all $K \in \mathcal{T}_h$.

Our starting point is the sesquilinear form $\widetilde{a}$ defined in (32). At the discrete level, we would like to localize the term $\langle (M-N)v, w\rangle_{V',V}$ at the boundary faces $F \in \mathcal{F}_h^\partial$. Therefore, we assume that there are boundary fields $\mathcal{M}$ and $\mathcal{N}$ in $L^\infty(\partial D;\mathbb{C}^{m\times m})$ such that

$$\langle Mv, w\rangle_{V',V} = (\mathcal{M}v, w)_{L^\partial}, \qquad \langle Nv, w\rangle_{V',V} = (\mathcal{N}v, w)_{L^\partial}, \tag{35}$$

for all $v, w \in V^s := H^s(D;\mathbb{C}^m)$ with $s > \frac{1}{2}$ and $L^\partial := L^2(\partial D;\mathbb{C}^m)$; whence,

$$\widetilde{a}(v,w) = (Av,w)_L + \frac{1}{2}((\mathcal{M}-\mathcal{N})v,w)_{L^\partial}, \qquad (v,w) \in V^s \times V^s. \tag{36}$$

The field $\mathcal{M}$ is such that $\Re((\mathcal{M}v,v)_{L^\partial}) \geq 0$, since the operator M is monotone. But it may occur that $\Re((\mathcal{M}v,v)_{L^\partial}) = 0$ (this happens for second-order PDEs in mixed form). To gain some control on the boundary values, we introduce an additional boundary penalty field $\mathcal{S}^\partial \in L^\infty(\partial D;\mathbb{C}^{m\times m})$, and we define the following sesquilinear form on $V^s \times V^s$:

$$\begin{aligned}\breve{a}(v,w) &:= \widetilde{a}(v,w) + (\mathcal{S}^\partial v,w)_{L^\partial} \\ &= (Av,w)_L + \frac{1}{2}((\mathcal{M}-\mathcal{N})v,w)_{L^\partial} + (\mathcal{S}^\partial v,w)_{L^\partial}.\end{aligned} \tag{37}$$

In what follows, we use a subscript F to denote the restriction of a boundary field to $F \in \mathcal{F}_h^\partial$, and we set $L(F) := L^2(F;\mathbb{C}^m)$. We define the local boundary semi-norm $|v|_{\mathcal{M}_F}^2 := (\mathcal{M}_F v, v)_{L(F)}$ and we set $\rho_F := \|\mathcal{M}_F\|_{L^\infty(F;\mathbb{C}^{m\times m})}$. We assume for simplicity that

$$\rho_F \leq c_{\mathcal{M}} \beta_{K_F}, \qquad \forall F \in \mathcal{F}_h^\partial, \tag{38}$$

where $K_F \in \mathcal{T}_h$ is the mesh element such that $F = \partial K_F \cap \partial D$. The design conditions on $\mathcal{S}^\partial$ are as follows: There is c, uniform with respect to h, such that the following holds for all $v, w \in L(F)$ and all $F \in \mathcal{F}_h^\partial$.

$$\mathcal{S}_F^\partial \text{ is Hermitian and positive semi-definite,} \tag{39a}$$

$$\ker(\mathcal{M}_F - \mathcal{N}_F) \subset \ker(\mathcal{S}_F^\partial), \tag{39b}$$

$$|v|_{\mathcal{S}_F^\partial} \leq c\, \rho_F^{\frac{1}{2}} \|v\|_{L(F)}, \tag{39c}$$

$$|((\mathcal{M}_F - \mathcal{N}_F)v,w)_{L(F)}| \leq c\,(|v|_{\mathcal{M}_F} + |v|_{\mathcal{S}_F^\partial})\rho_F^{\frac{1}{2}} \|w\|_{L(F)}, \tag{39d}$$

$$|((\mathcal{M}_F + \mathcal{N}_F)v,w)_{L(F)}| \leq c\, \rho_F^{\frac{1}{2}} \|v\|_{L(F)} (|w|_{\mathcal{M}_F} + |w|_{\mathcal{S}_F^\partial}). \tag{39e}$$

The assumption (39a) implies that the local boundary semi-norm $|y|_{\mathcal{S}_F^\partial}^2 := (\mathcal{S}_F^\partial v, v)_{L(F)}$ is well-defined and that $(\mathcal{S}_F^\partial v, w)_{L(F)} \leq |v|_{\mathcal{S}_F^\partial} |w|_{\mathcal{S}_F^\partial}$. The assumption (39b) is tailored to ensure exact consistency. The other assumptions (39c)-(39d)-(39e) are stability properties. Note that (39d)-(39e) turn out to be equivalent; both properties are presented since they are useful in the analysis.

Example 4.2 (Advection-reaction)**.** Since $\mathcal{M}_F = |\boldsymbol{\beta}\cdot\boldsymbol{n}_F|$ for all $F \in \mathcal{F}_h^\partial$, we can take $\mathcal{S}_F^\partial = 0$. The properties (39a), (39b), and (39c) are obvious, and (39d) results from the Cauchy–Schwarz inequality since $\frac{1}{2}\int_F (|\boldsymbol{\beta}\cdot\boldsymbol{n}_F| - \boldsymbol{\beta}\cdot\boldsymbol{n}_F) vw \,\mathrm{d}s \leq \||\boldsymbol{\beta}\cdot\boldsymbol{n}_F|^{\frac{1}{2}} v\|_{L^2(F)} \rho_F^{\frac{1}{2}} \|w\|_{L^2(F)}$. □

Example 4.3 (Maxwell)**.** Consider the boundary condition $\boldsymbol{H}\times\boldsymbol{n} = \boldsymbol{0}$ on ∂D. Recalling the matrix $\mathbb{T} \in \mathbb{R}^{3\times 3}$ from §1.3, the properties (39) are satisfied by taking $\mathcal{S}_F^\partial = \begin{pmatrix} \mathbb{O}_3 & \mathbb{O}_3 \\ \mathbb{O}_3 & \alpha \mathbb{T}^\intercal \mathbb{T} \end{pmatrix}$, for all $F \in \mathcal{F}_h^\partial$, with a parameter $\alpha > 0$. This means that the tangential component of $\boldsymbol{H}$ is penalized at the boundary. □

4.3. Galerkin/Least-Squares stabilization with boundary penalty

We define the following discrete sesquilinear form on $V_h \times V_h$:

$$\check{a}_h(v_h, w_h) = \widetilde{a}(v_h, w_h) + (Av_h, \tau A w_h)_L, \tag{40}$$

that is to say $\check{a}_h(v_h, w_h) = (Av_h, w_h)_L + \frac{1}{2}((\mathcal{M} - \mathcal{N})v_h, w_h)_{L^\partial} + (\mathcal{S}^\partial v_h, w_h)_{L^\partial} + (Av_h, \tau A w_h)_L$. We consider the following discrete problem:

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ \check{a}_h(u_h, w_h) = (f, w_h + \tau A w_h)_L, \quad \forall w_h \in V_h. \end{cases} \tag{41}$$

Let us set $V_\flat = V^s + V_h$. Notice that $V_\flat = V^s$ since the approximation is H^1-conforming. We extend the sesquilinear form $\check{a}_h$ to $V^s \times V_h$, and we equip the space V^s with the following norms:

$$\|v\|_{V_\flat}^2 := \mu_0 \|v\|_L^2 + \frac{1}{2}|v|_{\mathcal{M}}^2 + |v|_{\mathcal{S}^\partial}^2 + \|\tau^{\frac{1}{2}} A v\|_L^2, \tag{42a}$$

$$\|v\|_{V_{\flat\sharp}}^2 := \|v\|_{V_\flat}^2 + \|\tau^{-\frac{1}{2}} v\|_L^2 + \|\rho^{\frac{1}{2}} v\|_{L^\partial}^2, \tag{42b}$$

with boundary semi-norms $|v|_{\mathcal{M}}^2 := \Re((\mathcal{M}v, v)_{L^\partial})$ and $|v|_{\mathcal{S}^\partial}^2 := \Re((\mathcal{S}^\partial v, v)_{L^\partial})$, and $\rho \in L^\infty(\partial D)$ is defined by $\rho_{|F} := \rho_F$ for all $F \in \mathcal{F}_h^\partial$.

Theorem 4.4 (Convergence)**.** *(i) The discrete sesquilinear form $\check{a}_h$ satisfies $\Re(\check{a}_h(v_h, v_h)) \geq \|v_h\|_{V_\flat}^2$, for all $v_h \in V_h$. Consequently, the discrete problem* (41) *is well-posed. (ii) Assume that the exact solution u is in V^s. Then, the discrete problem* (41) *is exactly consistent. (iii) There is c, uniform with respect to h, such that $|\check{a}_h(v, w_h)| \leq |\check{a}(v, w_h)| + |(Av, \tau A w_h)_L| \leq c\,\|v\|_{V_{\flat\sharp}} \|w_h\|_{V_\flat}$ for all $(v, w_h) \in V^s \times V_h$. (iv) Let u and u_h be the unique solutions to* (19) *and* (41)*, respectively. Then, there is c, uniform with respect to h, such that*

$$\|u - u_h\|_{V_\flat} \leq c \inf_{v_h \in V_h} \|u - v_h\|_{V_{\flat\sharp}}. \tag{43}$$

Moreover, $\|u - u_h\|_{V_\flat}^2 \leq c \sum_{K \in \mathcal{T}_h} \max(\beta_K, \mu_0 h_K) h_K^{2r+1} |u|_{H^{1+r}(D_K;\mathbb{C}^m)}^2$ if $u \in H^{1+r}(D; \mathbb{C}^m)$, $r \in [0, k]$. This implies that $\|u - u_h\|_{V_\flat} \leq c \phi_D^{\frac{1}{2}} h^{r+\frac{1}{2}} |u|_{H^{1+r}(D;\mathbb{C}^m)}$.

5. Fluctuation-based stabilization

This section presents a unified analysis of various techniques for the approximation of first-order PDEs using H^1-conforming finite elements. The gradient of a function in an H^1-conforming space generally exhibits jumps across the mesh interfaces. This means that only one part of the gradient can be controlled by test functions from this space; the remainder, which can be viewed as a fluctuation, needs to be controlled by some stabilization mechanism. Three stabilization techniques are considered herein: the Continuous Interior Penalty (CIP), the Local Projection Stabilization (LPS), and the Subgrid Viscosity (SGV). CIP penalizes the jump of the gradient across the mesh interfaces. LPS and SGV

are both based on a two-scale decomposition of the discrete space consisting of a sum of resolved scales and fluctuations. LPS penalizes the fluctuations of the gradient, whereas SGV penalizes the gradient of the fluctuations. Throughout this section, the boundary conditions are enforced weakly by the boundary penalty technique introduced in §4.2.

5.1. Abstract theory for fluctuation-based stabilization

Let us consider the finite element setting introduced in §4.2. Let β_K and τ_K as defined in (25) and (26). Recall that β_K is a local velocity scale and τ_K is local time scale. We define the global quantity $\beta_D = \max_{K\in\mathcal{T}_h}\beta_K$, and we introduce a second local weighting parameter $\check{\tau}_K$ such that

$$\min(\beta_D^{-1}h_K, \mu_0^{-1}) \le \check{\tau}_K \le \tau_K, \qquad \forall K \in \mathcal{T}_h. \tag{44}$$

We will take $\check{\tau}_K = \min(\beta_D^{-1}h_K, \mu_0^{-1})$ for the CIP stabilization and $\check{\tau}_K = \tau_K$ for the LPS and SGV stabilizations. With a slight abuse of notation, we define the piecewise constant function $\check{\tau} : D \to \mathbb{R}$ such that $\check{\tau}_{|K} = \check{\tau}_K$ for all $K \in \mathcal{T}_h$; the piecewise constant function $\tau : D \to \mathbb{R}$ is defined similarly.

We additionally assume that all the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ are piecewise Lipschitz on a partition of D and that the meshes are compatible with this partition, implying that the fields $\{\mathcal{A}^k_{|K}\}_{k\in\{1:d\}}$ are Lipschitz for all $K \in \mathcal{T}_h$. We denote by $L_\mathcal{A}$ the largest Lipschitz constant of these fields. To simplify the tracking of the model parameters in the analysis, we assume that

$$\max(\|\mathcal{K}\|_{L^\infty(D;\mathbb{C}^{m\times m})}, \|\mathcal{X}\|_{L^\infty(D;\mathbb{C}^{m\times m})}, L_\mathcal{A}) \le c_{\mathcal{K},\mathcal{X},\mathcal{A}}\mu_0, \tag{45}$$

and we hide the non-dimensional factor $c_{\mathcal{K},\mathcal{X},\mathcal{A}}$ in the generic constant c.

The boundary conditions are enforced by using the boundary penalty method from §4.2, i.e., we assume that there is $\mathcal{S}^\partial \in L^\infty(\partial D;\mathbb{C}^{m\times m})$ satisfying (39) for any boundary face $F \in \mathcal{F}_h^\partial$, with $\rho_F = \|\mathcal{M}_F\|_{L^\infty(F;\mathbb{C}^{m\times m})}$. We assume that there is a uniform constant $c_\mathcal{M}$ such that $\rho_F \le c_\mathcal{M}\beta_{K_F}$ for all $F \in \mathcal{F}_h^\partial$ with $F = \partial K_F \cap \partial D$, see (38); we will hide the non-dimensional factor $c_\mathcal{M}$ in the generic constant c. Our starting point is the following sesquilinear form, see (37):

$$\check{a}(v,w) = (Av,w)_L + \frac{1}{2}((\mathcal{M}-\mathcal{N})v,w)_{L^\partial} + (\mathcal{S}^\partial v,w)_{L^\partial}, \quad \forall (v,w) \in V^s\times V^s. \tag{46}$$

The main idea is to augment the sesquilinear form $\check{a}$ with a stabilization sesquilinear form s_h and to consider the following discrete problem:

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ a_h(u_h,w_h) = (f,w_h)_L, \quad \forall w_h \in V_h, \end{cases} \tag{47}$$

with

$$a_h(v_h,w_h) := \check{a}(v_h,w_h) + s_h(v_h,w_h). \tag{48}$$

To stay somewhat general, we only require that s_h be defined on $V_h\times V_h$. Loosely speaking, the purpose of s_h is to control the difference between A_1v_h and a suitable representative of A_1v_h in V_h. We consider the following design requirements on the bilinear form s_h, where $c_1, c_2, c_3 > 0$ are uniform with respect to h:

(i) s_h is Hermitian positive semi-definite and satisfies $|v_h|_{\mathcal{S}} := s_h(v_h, v_h)^{\frac{1}{2}} \leq c_1 \|\check{\tau}^{-\frac{1}{2}} v_h\|_L$ for all $v_h \in V_h$.

(ii) There exists a linear map $\mathcal{J}_h : V_h \to V_h$ such that, for all $v_h \in V_h$,

$$c_2 \|\check{\tau}^{-\frac{1}{2}} \mathcal{J}_h(v_h)\|_L^2 \leq \|\check{\tau}^{\frac{1}{2}} A_1 v_h\|_L^2 + \mu_0 \|v_h\|_L^2 + |v_h|_{\mathcal{S}}^2, \tag{49a}$$

$$c_2 \|\check{\tau}^{\frac{1}{2}} A_1 v_h\|_L^2 \leq \Re((A_1 v_h, \mathcal{J}_h(v_h))_L) + \mu_0 \|v_h\|_L^2 + |v_h|_{\mathcal{S}}^2. \tag{49b}$$

(iii) $|\mathcal{I}_h(v)|_{\mathcal{S}} \leq c_3 \left(\sum_{K \in \mathcal{T}_h} (\check{\tau}_K^{-1} h_K) h_K^{2r+1} |v|^2_{H^{1+r}(D_K;\mathbb{C}^m)} \right)^{\frac{1}{2}}$ for all $r \in [0, k]$ and all $v \in H^{1+r}(D; \mathbb{C}^m)$ with $\mathcal{I}_h$ satisfying (34).

The error analysis is done in the spirit of Strang's First Lemma. This approach is the most general since it does not require that s_h be extended beyond $V_h \times V_h$. We consider the space $V_\flat = V^s + V_h$; note that $V_\flat = V^s$ since V_h is H^1-conforming. We define the following norms on V^s:

$$\|v\|_{V_\flat}^2 := \mu_0 \|v\|_L^2 + \frac{1}{2} |v|_{\mathcal{M}}^2 + |v|_{\mathcal{S}^\partial}^2 + \|\check{\tau}^{\frac{1}{2}} A_1 v\|_L^2, \tag{50a}$$

$$\|v\|_{V_{\flat\sharp}}^2 := \|v\|_{V_\flat}^2 + \|\check{\tau}^{-\frac{1}{2}} v\|_L^2 + \|\rho^{\frac{1}{2}} v\|_{L^\partial}^2. \tag{50b}$$

The first norm is used to establish the inf-sup stability of $\check{a}$ on $V_h \times V_h$ (and well-posedness) and the second one to prove the boundedness of of $\check{a}$ on $V^s \times V_h$. Up to the change of τ by $\check{\tau}$, these norms are the same as those used in §4.3 for the GaLS stabilization with boundary penalty.

Theorem 5.1 (Convergence)**.** *(i) Under the design conditions* (i)-(ii)-(iii) *for s_h, there is $\alpha > 0$, uniform with respect to h, such that the following holds:*

$$\alpha(\|v_h\|_{V_\flat} + |v_h|_{\mathcal{S}}) \leq \sup_{w_h \in V_h} \frac{\Re(a_h(v_h, w_h))}{\|w_h\|_{V_\flat} + |w_h|_{\mathcal{S}}}, \qquad \forall v_h \in V_h. \tag{51}$$

Consequently, the discrete problem (47) *is well-posed. (ii) There is c, uniform with respect to h, such that $|\check{a}(v, w_h)| \leq c \|v\|_{V_{\flat\sharp}} \|w_h\|_{V_\flat}$ holds for all $(v, w_h) \in V^s \times V_h$. (iii) Let u be the unique solution to* (19) *and let u_h be the unique solution to* (47) *with s_h satisfying the design conditions* (i)-(ii)-(iii) *above. There is c, uniform with respect to h, such that*

$$\|u - u_h\|_{V_\flat} \leq c \inf_{v_h \in V_h} \left(\|u - v_h\|_{V_{\flat\sharp}} + |v_h|_{\mathcal{S}} \right). \tag{52}$$

Moreover, $\|u - u_h\|_{V_\flat}^2 \leq c \sum_{K \in \mathcal{T}_h} (\check{\tau}_K^{-1} h_K) h_K^{2r+1} |u|^2_{H^{1+r}(D_K;\mathbb{C}^m)}$ if $u \in H^{1+r}(D; \mathbb{C}^m)$, $r \in [0, k]$ (note that $\max(\beta_K, \mu_0 h_K) \leq \check{\tau}_K^{-1} h_K \leq \max(\beta_D, \mu_0 h_K)$).

In the next section we show how the above theory can be used to analyze the stability and convergence properties of the Continuous Interior Penalty (CIP), the Local Projection Stabilization (LPS), and the Subgrid Viscosity (SGV).

5.2. Continuous Interior Penalty

The key idea in CIP stabilization (also termed edge stabilization in the literature) is to penalize the jump of $A_1 v_h$ across the mesh interfaces. This idea has been introduced in Burman [6], Burman and Hansbo [9]. We refer to [7, 8] for the *hp* analysis and extensions to Friedrichs' systems, and we refer to [15] for extensions in the context of nonlinear conservation laws.

We set $\check{\tau}_K := \min(\beta_D^{-1} h_K, \mu_0^{-1})$ for all $K \in \mathcal{T}_h$. Let us take $V_h = P_k^{\mathrm{g}}(\mathcal{T}_h; \mathbb{C}^m)$. Let $\mathcal{J}_h^{\mathrm{g,av}}$ be the nodal averaging operator mapping onto $P_k^{\mathrm{g}}(\mathcal{T}_h; \mathbb{C}^m)$ defined and analyzed in [16], and let $\phi \in P_1^{\mathrm{g}}(\mathcal{T}_h; \mathbb{R})$ be defined by $\phi(\boldsymbol{z}) = \mathrm{card}(\mathcal{T}_{\boldsymbol{z}})^{-1} \sum_{K \in \mathcal{T}_{\boldsymbol{z}}} \check{\tau}_K$ with $\mathcal{T}_{\boldsymbol{z}} := \{K \in \mathcal{T}_h \mid \boldsymbol{z} \in K\}$ for any mesh vertex $\boldsymbol{z}$.

Lemma 5.2. *Define* $(\underline{A}_1 v_h)_{|K} := \sum_{k \in \{1:d\}} \underline{\mathcal{A}}_K^k \partial_k v_{h|K}$ *for all* $K \in \mathcal{T}_h$ *and all* $v_h \in V_h$, *where* $\underline{\mathcal{A}}_K^k := \frac{1}{|K|} \int_K \mathcal{A}^k \,\mathrm{d}x$. *Let* $\check{\tau}_F := \max(\check{\tau}_{K_l}, \check{\tau}_{K_r})$ *and* $\beta_F := \max(\beta_{K_l}, \beta_{K_r})$ *for all* $F = \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$. *Then the sesquilinear forms*

$$s_h^{\mathrm{CIP}}(v_h, w_h) = \sum_{F \in \mathcal{F}_h^\circ} \check{\tau}_F h_F (\llbracket A_1 v_h \rrbracket_F, \llbracket A_1 w_h \rrbracket_F)_{L(F)}, \tag{53a}$$

$$s_h^{\mathrm{CIP}}(v_h, w_h) = \sum_{F \in \mathcal{F}_h^\circ} \check{\tau}_F h_F (\llbracket \underline{A}_1 v_h \rrbracket_F, \llbracket \underline{A}_1 w_h \rrbracket_F)_{L(F)}, \tag{53b}$$

$$s_h^{\mathrm{CIP}}(v_h, w_h) = \sum_{F \in \mathcal{F}_h^\circ} \beta_F h_F^2 (\llbracket \nabla v_h \rrbracket_F, \llbracket \nabla w_h \rrbracket_F)_{\boldsymbol{L}(F)}, \tag{53c}$$

all satisfy the conditions (i)-(ii) *with* $\mathcal{J}_h(v_h) = \mathcal{J}_h^{\mathrm{g,av}}(\phi \underline{A}_1 v_h)$, *and the condition* (iii) *for* $r \geq 1$.

Remark 5.3 (Time-dependent case)**.** The choice (53c) is interesting for time-dependent fields $\mathcal{A}^k$ since the matrix associated with (53c) can then be assembled only once, which is not the case for (53a)-(53b). Note that in (53c), only the normal component of the gradient can actually jump across F since functions in V_h are continuous. □

5.3. Two-scale stabilization: Local Projection and Subgrid Viscosity

We present in this section two closely related stabilization techniques known in the literature as Local Projection Stabilization (LPS) and Subgrid Viscosity (SGV). The SGV technique has been introduced in Guermond [19, 20, 21] for monotone operators and semi-groups. The LPS technique has been introduced in Becker and Braack [1], Braack and Burman [2] for Stokes and convection-diffusion equations; see also Matthies et al. [25, 26]. LPS and SGV both rely on a two-scale decomposition of the discrete space V_h, leading to the notions of resolved and fluctuating (or subgrid) scales. Both stabilization techniques introduce a least-squares penalty: LPS penalizes the fluctuation of the gradient and SGV penalizes the gradient of the fluctuation. The notion of scale separation and subgrid scale dissipation is similar in spirit to the spectral viscosity technique introduced by Tadmor [28] to approximate nonlinear conservation equations by means of spectral methods. This notion is also found in the Orthogonal Subscale Stabilization technique of Codina [10].

5.3.1. The two-scale decomposition

The starting point is a two-scale decomposition of V_h into the form

$$V_h = R_h + B_h, \tag{54}$$

where the sum is not necessarily direct. The discrete space R_h is viewed as the space of the resolved scales, and B_h is viewed as the space of the fluctuating (or subgrid) scales. It is important to realize that the degrees of freedom attached to B_h only serve to achieve stability, and that the approximation error is controlled by the best approximation in the space of the resolved scales R_h (and not in the full space V_h). We assume the following local approximation property in R_h: There is a quasi-interpolation operator $\mathcal{I}_h^R : V \to R_h$ and a constant c, uniform with respect to h, such that

$$\|v - \mathcal{I}_h^R(v)\|_{L(K)} + h\|\nabla(v - \mathcal{I}_h^R(v))\|_{\boldsymbol{L}(K)} \le c\, h^{1+r}|v|_{H^{1+r}(D_K;\mathbb{C}^m)}, \tag{55}$$

for all $r \in [0,k]$, all $v \in H^{1+r}(D;\mathbb{C}^m)$, and all $K \in \mathcal{T}_h$.

Since functions in R_h are continuous, piecewise polynomials, the components of their gradients belong to a broken finite element space $G_h = \bigoplus_{K\in\mathcal{T}_h} G_K$, where functions in G_K are supported in K, i.e., $\partial_i r_h \in G_h$ for all $r_h \in R_h$ and all $i \in \{1{:}d\}$. We assume that the space of the fluctuating scales can also be localized in the form $B_h = \bigoplus_{K\in\mathcal{T}_h} B_K$, where the functions in B_K are supported in K (one may think of members of B_K as bubble-type functions, see the examples below). We define the local L-orthogonal projections $\pi_K^B : L(K) \to B_K$ and $\pi_K^G : L(K) \to G_K$ for all $K \in \mathcal{T}_h$ and the global counterparts $\pi_h^B : L \to B_h$ and $\pi_h^G : L \to G_h$ such that $\pi_{h|K}^B = \pi_K^B$ and $\pi_{h|K}^G = \pi_K^G$.

The key assumption linking the local gradient space G_K to the local fluctuation space B_K is the following inf-sup condition introduced in [19, 20] (see also [25]): There is $\gamma > 0$, uniform with respect to h, such that, for all $K \in \mathcal{T}_h$,

$$\inf_{g\in G_K}\sup_{b\in B_K}\frac{\Re(\int_K b^{\mathsf{H}} g\,\mathrm{d}x)}{\|g\|_{L(K)}\|b\|_{L(K)}} \ge \gamma, \tag{56}$$

or, equivalently, $\gamma\|g\|_{L(K)} \le \|\pi_K^B g\|_{L(K)}$ for all $g \in G_K$. In what follows, we consider the local weighting parameter $\check{\tau}_K = \tau_K = \min(\beta_K h_K^{-1}, \mu_0^{-1})$ for all $K \in \mathcal{T}_h$.

We now describe three constructions of H^1-conforming finite element spaces of degree $k \ge 1$ which all satisfy the above assumptions. (1) In the first example, the space of the resolved scales is defined by $R_h = P_k^{\mathrm{g}}(\mathcal{T}_h;\mathbb{C}^m)$, the H^1-conforming finite element space based on $\mathcal{T}_h$, so that $G_h = P_{k-1}^{\mathrm{b}}(\mathcal{T}_h;\mathbb{C}^m)$ and G_K is composed of $\mathbb{C}^m$-valued polynomials of degree at most $(k-1)$ on affine meshes. Following [19] for $k \in \{1,2\}$ and [25] for all $k \ge 1$, we take $B_K = b_K G_K$ where b_K is the $H_0^1(K)$-bubble function proportional to the product of the $(d+1)$ barycentric coordinates over K; see the panels in the upper row in Figure 1. (2) Instead of working with bubble functions, one can use hierarchical meshes [19, 25]. In this case, the construction starts from the mesh defining the space of the resolved scales, say $\check{\mathcal{T}}_h$. Assume for simplicity that $\check{\mathcal{T}}_h$

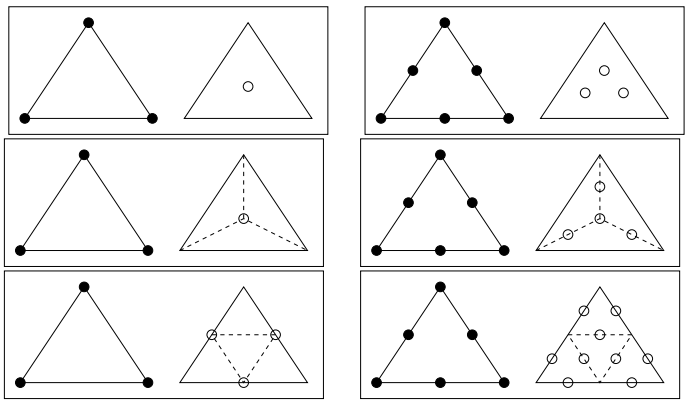

Figure 1: Examples of two-scale finite elements. In each panel, the resolved scales are on the left and the fluctuating scales are on the right. The resolved scales are either $\mathbb{P}_1$ (left column) or $\mathbb{P}_2$ (right column) Lagrange elements. The upper panels illustrate the use of a standard bubble function to build the fluctuating scales; the central and lower panels illustrate the use of piecewise polynomial bubble functions on a submesh with the same size (central panel) or half the size (bottom panel) as that of the resolved scales space.

is composed of simplices, then the mesh $\mathcal{T}_h$ defining V_h is built by barycentric refinement, i.e., for any $K \in \check{\mathcal{T}}_h$, $(d+1)$ new simplices are created by joining the barycenter of K to its $(d+1)$ vertices. Then we take $V_h = P_k^{\mathrm{g}}(\mathcal{T}_h;\mathbb{C}^m)$ and $R_h = P_k^{\mathrm{g}}(\check{\mathcal{T}}_h;\mathbb{C}^m)$, so that $G_h = P_{k-1}^{\mathrm{b}}(\check{\mathcal{T}}_h;\mathbb{C}^m)$, see the panels in the second row in Figure 1. For any $K \in \check{\mathcal{T}}_h$, choose $g := \dim(G_K)$ shape functions of V_h with support in K, say $\varphi_1^K, \dots, \varphi_g^K$ and set $B_K = \mathrm{span}\{\varphi_1^K, \dots, \varphi_g^K\}$. The practical advantage of this construction is that V_h is a standard finite element space. (3) Finally, we mention the two-scale decomposition considered in [19] for $k \in \{1,2\}$ which also offers the advantage of V_h being a standard finite element space; a schematic representation of the scale decomposition is shown in the panels in the last row in Figure 1. The analysis (not considered herein) is somewhat more involved since the fluctuating scales are represented by functions possibly supported on two adjacent mesh cells.

5.3.2. Local Projection Stabilization

Lemma 5.4. *Assume that* (56) *holds. Let* $\underline{A}_1 v_h$ *be defined as in Lemma 5.2 and* $\beta : D \to \mathbb{R}$ *be such that* $\beta_{|K} := \beta_K$ *for all* $K \in \mathcal{T}_h$*. Define the fluctuation operator* $\kappa_h^G = I_L - \pi_h^G$*, where* I_L *is the identity operator in* L*. Then, the sesquilinear forms*

$$s_h^{\mathrm{LPS}}(v_h, w_h) = (\check{\tau}\kappa_h^G(\underline{A}_1 v_h), \kappa_h^G(\underline{A}_1 w_h))_L, \tag{57a}$$

$$s_h^{\mathrm{LPS}}(v_h, w_h) = (\beta^2\check{\tau}\kappa_h^G(\nabla v_h), \kappa_h^G(\nabla w_h))_{\boldsymbol{L}}, \tag{57b}$$

both satisfy the assumptions (i)-(ii)-(iii) *with* $\mathcal{J}_h(v_h) = \check{\tau}\pi_h^B\pi_h^G(A_1 v_h)$.

Remark 5.5 (Use of $\kappa_h^G(A_1 v_h)$)**.** When the fields $\mathcal{A}^k$ are not piecewise constant, setting $s_h^{\mathrm{LPS}}(v_h, w_h) = (\check{\tau}\kappa_h^G(A_1 v_h), \kappa_h^G(A_1 w_h))_L$ is somewhat delicate

since $|\mathcal{I}_h^R(u)|_{\mathcal{S}}$ no longer vanishes. Bounding this quantity requires strong regularity assumptions on the fields $\mathcal{A}^k$. □

5.3.3. Subgrid Viscosity

In the SGV method, the two-scale decomposition of V_h is assumed to be direct and L-stable, i.e., it is assumed that there is $\gamma_R > 0$, uniform with respect to h, such that

$$V_h = R_h \oplus B_h, \qquad \gamma_R \|\pi_h^R v_h\|_L \le \|v_h\|_L, \qquad \forall v_h \in V_h. \tag{58}$$

Letting $\pi_h^R : V_h \to R_h$ be the oblique projector based on (58), we define the fluctuation operator $\kappa_h^R := I_{V_h} - \pi_h^R$, where I_{V_h} the identity in V_h. Just as for LPS stabilization, we can choose $R_h = P_k^{\mathrm{g}}(\mathcal{T}_h)$. Then, G_h is the broken finite element space $P_{k-1}^{\mathrm{b}}(\mathcal{T}_h)$, i.e., $G_K = \mathbb{P}_{k-1,d}$ on simplicial affine meshes (d-variate polynomials of order at most $k-1$). The simple choice $B_K = b_K G_K$ is only possible for $k \le d$, since otherwise the decomposition (58) is no longer direct. For $k \ge d+1$, a simple possibility to get around this technicality is to set $B_K = b_K^\alpha G_K$ with α equal to $\frac{k+1}{d+1}$ or to the smallest integer larger than $\frac{k}{d+1}$, see also [19, Prop. 4.1].

Lemma 5.6. *Assume that* (56) *holds. Let $\beta : D \to \mathbb{R}$ be such that $\beta_{|K} := \beta_K$ for all $K \in \mathcal{T}_h$. Then the sesquilinear forms*

$$s_h^{\mathrm{SGV}}(v_h, w_h) = (\check{\tau} A_1(\kappa_h^R v_h), A_1(\kappa_h^R w_h))_L, \tag{59a}$$

$$s_h^{\mathrm{SGV}}(v_h, w_h) = (\check{\tau} \underline{A}_1(\kappa_h^R v_h), \underline{A}_1(\kappa_h^R w_h))_L, \tag{59b}$$

$$s_h^{\mathrm{SGV}}(v_h, w_h) = (\beta^2 \check{\tau} \nabla(\kappa_h^R v_h), \nabla(\kappa_h^R w_h))_{\boldsymbol{L}}, \tag{59c}$$

all satisfy the assumptions (i)-(ii)-(iii) *with $\mathcal{J}_h(v_h) = \check{\tau} \pi_h^B \underline{A}_1(\pi_h^R(v_h))$.*

References

[1] R. Becker and M. Braack. A finite element pressure gradient stabilization for the Stokes equations based on local projections. *Calcolo*, 38(4):173–199, 2001.

[2] M. Braack and E. Burman. Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method. *SIAM J. Numer. Anal.*, 43(6):2544–2566, 2006.

[3] J. H. Bramble and A. H. Schatz. Rayleigh-Ritz-Galerkin-methods for Dirichlet's problem using subspaces without boundary conditions. *Comm. Pure Appl. Math.*, 23:653–675, 1970.

[4] J. H. Bramble and A. H. Schatz. Least squares for $2m$th order elliptic boundary-value problems. *Math. Comp.*, 25:1–32, 1971.

[5] A. Brooks and T. Hughes. Streamline Upwind/Petrov–Galerkin formulations for convective dominated flows with particular emphasis on the incompressible Navier–Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 32:199–259, 1982.

[6] E. Burman. A unified analysis for conforming and nonconforming stabilized finite element methods using interior penalty. *SIAM J. Numer. Anal.*, 43(5):2012–2033 (electronic), 2005.

[7] E. Burman and A. Ern. Continuous interior penalty hp-finite element methods for advection and advection-diffusion equations. *Math. Comp.*, 76(259):1119–1140, 2007.

[8] E. Burman and A. Ern. A continuous finite element method with face penalty to approximate Friedrichs' systems. *M2AN Math. Model. Numer. Anal.*, 41(1):55–76, 2007.

[9] E. Burman and P. Hansbo. Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193(15-16):1437–1453, 2004.

[10] R. Codina. Stabilized finite element approximation of transient incompressible flows using orthogonal subscales. *Comput. Methods Appl. Mech. Engrg.*, 191(39-40):4295–4321, 2002.

[11] A. Džiškariani. The least square and Bubnov-Galerkin methods. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 8:1110–1116, 1968.
[12] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.
[13] A. Ern and J.-L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. I. General theory. *SIAM J. Numer. Anal.*, 44(2):753–778, 2006.
[14] A. Ern and J.-L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. II. Second-order elliptic PDEs. *SIAM J. Numer. Anal.*, 44(6):2363–2388, 2006.
[15] A. Ern and J.-L. Guermond. Weighting the edge stabilization. *SIAM J. Numer. Anal.*, 51(3): 1655–1677, 2013.
[16] A. Ern and J.-L. Guermond. Finite element quasi-interpolation and best approximation. *ESAIM: Math. Model. Numer. Anal.*, 2016. to appear, preprint available at `http://arxiv.org/abs/1505.06931`.
[17] A. Ern, J.-L. Guermond, and G. Caplain. An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrichs' systems. *Comm. Partial Differ. Eq.*, 32:317–341, 2007.
[18] K. Friedrichs. Symmetric positive linear differential equations. *Comm. Pure Appl. Math.*, 11: 333–418, 1958.
[19] J.-L. Guermond. Stabilization of Galerkin approximations of transport equations by subgrid modeling. *M2AN Math. Model. Numer. Anal.*, 33(6):1293–1316, 1999.
[20] J.-L. Guermond. Subgrid stabilization of Galerkin approximations of linear monotone operators. *IMA J. Numer. Anal.*, 21:165–197, 2001.
[21] J.-L. Guermond. Subgrid stabilization of Galerkin approximations of linear contraction semigroups of class C^0 in Hilbert spaces. *Numer. Methods Partial Differential Equations*, 17(1): 1–25, 2001.
[22] T. Hughes, L. Franca, and G. Hulbert. A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/Least-Squares method for advection-diffusive equations. *Comput. Methods Appl. Mech. Engrg.*, 73:173–189, 1989.
[23] C. Johnson, U. Nävert, and J. Pitkäranta. Finite element methods for linear hyperbolic equations. *Comput. Methods Appl. Mech. Engrg.*, 45:285–312, 1984.
[24] A. Lučka. The rate of convergence to zero of the residual and the error for the Bubnov-Galerkin method and the method of least squares. In *Proc. Sem. Differential and Integral Equations, No. I (Russian)*, pages 113–122. Akad. Nauk Ukrain. SSR Inst. Mat., Kiev, Ukraine, 1969.
[25] G. Matthies, P. Skrzypacz, and L. Tobiska. A unified convergence analysis for local projection stabilisations applied to the Oseen problem. *M2AN Math. Model. Numer. Anal.*, 41(4):713–742, 2007.
[26] G. Matthies, P. Skrzypacz, and L. Tobiska. Stabilization of local projection type applied to convection-diffusion problems with mixed boundary conditions. *Electron. Trans. Numer. Anal.*, 32:90–105, 2008.
[27] J. Rauch. Boundary value problems with nonuniformly characteristic boundary. *J. Math. Pures Appl. (9)*, 73(4):347–353, 1994.
[28] E. Tadmor. Convergence of spectral methods for nonlinear conservation laws. *SIAM J. Numer. Anal.*, 26(1):30–44, 1989.