Numerical Analysis of Partial Differential Equations

Alfio Quarteroni

MOX, Dipartimento di Matematica
Politecnico di Milano

Lecture Notes
A.Y. 2022-2023

Advection–diffusion–reaction (ADR) equations

Cfr [Q], Chap. 13

We consider the problem $\mathcal{L}u = f$ in Ω, $u = 0$ on $\partial\Omega$, where:

1. $\mathcal{L}u = -\operatorname{div}(\mu\nabla u + \mathbf{b}u) + \sigma u$ (conservative form)
2. $\mathcal{L}u = -\operatorname{div}(\mu\nabla u) + \mathbf{b}\cdot\nabla u + \sigma u$ (non-conservative form)

Assumptions on coefficients as in Lecture 1 (see slide 8).

Weak formulation:

$$\begin{cases} \text{Find } u \in V = H^1_0(\Omega) \\ a(u,v) = F(v) \qquad \forall\, v \in V \end{cases} \tag{1}$$

$$F(v) = \int_\Omega f\, v$$

$$a(u,v) = \begin{cases} \int_\Omega (\mu\nabla u + \mathbf{b}u)\cdot\nabla v + \int_\Omega \sigma\, u\, v & \text{conservative} \\ \int_\Omega \mu\nabla u\cdot\nabla v + \int_\Omega \mathbf{b}\cdot\nabla u\, v + \int_\Omega \sigma\, u\, v & \text{non-conservative} \end{cases} \tag{2}$$

[Q] A. Quarteroni, *Numerical Models for Differential Problems*, 3rd Ed. , Springer, 2018

Coercivity

Sufficient conditions for coercivity:

1. Non-conservative case: $\sigma - \frac{1}{2}\,\mathrm{div}\,\mathbf{b} \geq 0$ in Ω
   (see Lecture 1, slides 22–23)
2. Conservative case: $\sigma + \frac{1}{2}\,\mathrm{div}\,\mathbf{b} \geq 0$ in Ω
   (prove it – similar proof)

In both cases: $a(u,u) \geq \mu_0 \|\nabla u\|^2 \quad \rightarrow \quad$ coercivity constant $\alpha \simeq \mu_0$

Continuity

In both cases, continuity constant: $M \simeq \|\mu\|_{L^\infty} + \|\mathbf{b}\|_{L^\infty} + \|\sigma\|_{L^2}$

(see Lecture 1, slide 21, for the non-conservative case. Similar proof for the conservative case – see book [Q])

[Q] A. Quarteroni, *Numerical Models for Differential Problems*, 3rd Ed. , Springer, 2018

FE error estimate

$$\|u - u_h\| \overset{\text{(Céa)}}{\leq} \frac{M}{\alpha} \inf_{v_h \in V_h} \|u - v_h\| \overset{\substack{\text{(interpolation}\\ \text{error estimate)}}}{\leq} C \frac{M}{\alpha} h^r |u|_{H^{r+1}(\Omega)} \qquad (3)$$

If convection dominated flow (or reaction dominated flow), then $M/\alpha \gg 1$:
$\rightarrow$ trade-off between M/α and h^r
$\rightarrow$ numerically prohibitive

$$\boxed{\mathbb{P}\text{e} = h\frac{M}{\alpha}}$$

Peclet number – "moral definition". More precise definition an a case-by-case mode.

$\rightarrow$ $\mathbb{P}$e should be less than 1 (for stability issues – see later).

Idea: stabilize the Galerkin method

1D case: Upwind method $\iff$ Artificial diffusion

2D case: Streamline diffusion (see lab)

$$+c(h)\int_\Omega \frac{1}{\|\mathbf{b}\|}(\mathbf{b}\cdot\nabla u_h)(\mathbf{b}\cdot\nabla v_h)$$

Artificial diffusion (diffuse everywhere)

$$+c(h)\int_\Omega \nabla u_h\cdot\nabla v_h$$

Not fully/strongly consistent!

$$\begin{cases} \text{Find } u_h \in V_h \\ a(u_h, v_h) + \mathscr{L}_h(u_h, f; v_h) = F(v_h) \qquad \forall\, v_h \in V_h \end{cases} \tag{4}$$

$\mathscr{L}_h$ suitably chosen – should be such that:

$$\boxed{\mathscr{L}_h(u, f; v_h) = 0 \qquad \forall\, v_h \in V_h}$$

$\rightarrow$ **strongly consistent** approximation of the original problem.

Idea: proportional to the residual

$$\mathscr{L}_h(u_h, f; v_h) = \sum_{K \in \mathcal{T}_h} \int_K (\mathcal{L} u_h - f)\, \tau_K \phi(v_h) \qquad \forall\, v_h \in V_h \tag{5}$$

τ_K: scaling factor. Typical choice:

$$\begin{aligned} \tau_K(\mathbf{x}) &= \delta \frac{h_K}{|\mathbf{b}(\mathbf{x})|} \qquad \forall\, \mathbf{x} \in K, K \in \mathcal{T}_h \\ h_K &= \mathrm{diam}(K) \end{aligned} \tag{6}$$

Many ways of choosing $\phi(v_h)$. Two remarkable choices:

1. $\phi(v_h) = \mathcal{L} v_h \rightarrow$ **GLS** – Galerkin least square method
2. $\phi(v_h) = \mathcal{L}_{\text{ss}} v_h \rightarrow$ **SUPG** – Streamline upwind Petrov-Galerkin method

Notation: $\mathcal{L} = \mathcal{L}_{\text{s}} + \mathcal{L}_{\text{ss}}$ (symmetric + skew-symmetric part of $\mathcal{L}$)

Definitions:

$$
\begin{aligned}
{}_{V'}\langle \mathcal{L}_{\text{s}} u, v \rangle_V &= {}_V\langle u, \mathcal{L}_{\text{s}} v \rangle_{V'} \qquad \forall\, u, v \in V \\
{}_{V'}\langle \mathcal{L}_{\text{ss}} u, v \rangle_V &= -{}_V\langle u, \mathcal{L}_{\text{ss}} v \rangle_{V'} \qquad \forall\, u, v \in V
\end{aligned}
$$

Remark: for matrices, $A = A_{\text{S}} + A_{\text{SS}}$, with:

$$
A_{\text{S}} = \frac{1}{2}(A + A^T), \qquad A_{\text{SS}} = \frac{1}{2}(A - A^T)
$$

Example (Non-conservative form)

$$\mathcal{L}^1 = -\mu\Delta u + \mathbf{b}\cdot\nabla u + \sigma u$$
$$= \underbrace{\left[-\mu\Delta u + \left(\sigma - \frac{1}{2}\operatorname{div}\mathbf{b}\right) u\right]}_{\mathcal{L}^1_{\mathrm{s}} u} + \underbrace{\left[\frac{1}{2}\left(\operatorname{div}(\mathbf{b}u) + \mathbf{b}\cdot\nabla u\right)\right]}_{\mathcal{L}^1_{\mathrm{ss}} u}$$

Indeed:

$$_{V'}\langle \mathcal{L}^1_{\mathrm{s}} u, v\rangle_V = \int_\Omega \mu\nabla u\cdot\nabla v + \left(\sigma - \frac{1}{2}\operatorname{div}\mathbf{b}\right) u\, v$$
$$= \int_\Omega \left[-\mu\Delta v + \left(\sigma - \frac{1}{2}\operatorname{div}\mathbf{b}\right) v\right] u = {}_V\langle u, \mathcal{L}^1_{\mathrm{s}} v\rangle_{V'}$$

$$_{V'}\langle \mathcal{L}^1_{\mathrm{ss}} u, v\rangle_V = \frac{1}{2}\int_\Omega \left(\operatorname{div}(\mathbf{b}u)v + (\mathbf{b}\cdot\nabla u)\, v\right)$$
$$= \frac{1}{2}\int_\Omega \left(-(\mathbf{b}u)\cdot\nabla v + (\mathbf{b}v)\cdot\nabla u\right)$$
$$= \frac{1}{2}\int_\Omega \left(-(\mathbf{b}\cdot\nabla v)u - \operatorname{div}(\mathbf{b}v)u\right) = -{}_V\langle u, \mathcal{L}^1_{\mathrm{ss}} v\rangle_{V'}$$

Example (Conservative form)

$$
\begin{aligned}
\mathcal{L}^2 &= -\mu\Delta u + \mathrm{div}(\mathbf{b}u) + \sigma u \\
&= \underbrace{\left[-\mu\Delta u + \left(\sigma + \frac{1}{2}\,\mathrm{div}\,\mathbf{b}\right) u\right]}_{\mathcal{L}^2_{\mathrm{s}} u} + \underbrace{\left[\frac{1}{2}\,(\mathrm{div}(\mathbf{b}u) + \mathbf{b}\cdot\nabla u)\right]}_{\mathcal{L}^2_{\mathrm{ss}} u}
\end{aligned}
$$

The proof is similar (do it yourself).

Remark

If $\operatorname{div}\mathbf{b} = 0$ (this happens, for instance, when $\mathbf{b}$ is constant), then the conservative and non-conservative forms coincide: $\mathcal{L}^1 = \mathcal{L}^2$. Indeed.

$$\operatorname{div}(\mathbf{b}u) = \mathbf{b} \cdot \nabla u$$

In this case:

$$\mathcal{L}_{\mathrm{s}} u = -\mu \Delta u + \sigma u, \qquad \mathcal{L}_{\mathrm{ss}} u = \mathbf{b} \cdot \nabla u$$

Indeed:

$$\begin{aligned} {}_{V'}\langle \mathcal{L}^1_{\mathrm{s}} u, v \rangle_V &= (\mu \nabla u, \nabla v) + (\sigma u, v) = {}_V\langle u, \mathcal{L}^1_{\mathrm{s}} v \rangle_{V'} \\ {}_{V'}\langle \mathcal{L}^1_{\mathrm{ss}} u, v \rangle_V &= (\mathbf{b} \cdot \nabla u, v) = (\nabla u, \mathbf{b} v) \\ &= -(u, \operatorname{div}(\mathbf{b} v)) = -(u, \mathbf{b} \cdot \nabla v) = -{}_V\langle u, \mathcal{L}^1_{\mathrm{ss}} v \rangle_{V'} \end{aligned}$$

Back to stabilized Galerkin.

Remark

Note that if $r = 1$, $\sigma = 0$ and $\operatorname{div} \mathbf{b} = 0$, the two methods SUPG and GLS coincide. Indeed, $-\Delta u_h|_K = 0$ on each $K \in \mathcal{T}_h$.

Problem in conservative form – GLS method

$$\begin{cases} \text{Find } u_h \in V_h \\ a(u_h, v_h) + \displaystyle\sum_{K \in \mathcal{T}_h} \int_K \mathcal{L} u_h \, \tau_K \, \mathcal{L} v_h = \\ \qquad \displaystyle\int_\Omega f \, v_h + \sum_{K \in \mathcal{T}_h} \int_K f \, \tau_K \, \mathcal{L} v_h \qquad \forall \, v_h \in V_h \end{cases} \tag{7}$$

which can be rewritten (with obvious meaning of notations) as:

$$\begin{cases} \text{Find } u_h \in V_h \\ a_h(u_h, v_h) = F_h(v_h) \qquad \forall \, v_h \in V_h \end{cases} \tag{8}$$

Stability analysis

Theorem

Consider the conservative case. Suppose that

$$\exists\, \gamma_0, \gamma_1 > 0\colon\ 0 < \gamma_0 \leq \gamma(\mathbf{x}) \leq \gamma_1 \tag{9}$$

then, for a suitable constant C independent of h, we have:

$$\|u_h\|^2_{GLS} \leq C\ \|f\|^2_{L^2(\Omega)}$$

($\|\cdot\|_{GLS}$ to be defined later).

Proof. Take $v_h = u_h$. We have:

$$a_h(u_h, u_h) = \int_\Omega \mu|\nabla u_h|^2 + \underbrace{\int_\Omega \mathrm{div}(\mathbf{b}u_h)u_h}_{=-\int_\Omega \mathbf{b}\cdot(u_h\nabla u_h)=-\frac{1}{2}\int_\Omega \mathbf{b}\cdot\nabla(u_h^2)=\frac{1}{2}\int_\Omega \mathrm{div}(\mathbf{b})u_h^2} + \int_\Omega \sigma u_h^2 + \sum_{K\in\mathcal{T}_h}\int_K \tau_K\,(\mathcal{L}u_h)^2$$

$$= \int_\Omega \mu|\nabla u_h|^2 + \int_\Omega \underbrace{\left(\sigma + \frac{1}{2}\,\mathrm{div}\,\mathbf{b}\right)}_{=:\gamma(\mathbf{x})} u_h^2 + \sum_{K\in\mathcal{T}_h}\int_K \tau_K\,(\mathcal{L}u_h)^2$$

$$=: \|u_h\|^2_{\mathrm{GLS}}$$

On the other hand:

$$|F_h(u_h)| \leq \left| \int_\Omega f\, u_h \right| + \left| \sum_{K \in \mathcal{T}_h} \int_K f\, \tau_K\, \mathcal{L} u_h \right|$$

Where:

$$\begin{aligned}
\left| \int_\Omega f\, u_h \right| &= \left| \int_\Omega \frac{1}{\sqrt{\gamma}} f\, \sqrt{\gamma} u_h \right| \\
&\overset{\text{(Cauchy–Schwarz)}}{\leq} \left\| \frac{1}{\sqrt{\gamma}} f \right\|_{L^2(\Omega)} \|\sqrt{\gamma} u_h\|_{L^2(\Omega)} \\
&\overset{\text{(Young*)}}{\leq} \left\| \frac{1}{\sqrt{\gamma}} f \right\|^2_{L^2(\Omega)} + \frac{1}{4} \|\sqrt{\gamma} u_h\|^2_{L^2(\Omega)}
\end{aligned}$$

* Young inequality: $AB \leq \epsilon A^2 + \frac{1}{4\epsilon} B^2 \quad \forall A, B \in \mathbb{R}, \epsilon > 0$

And where:

$$\left|\sum_{K\in\mathcal{T}_h}\int_K f\,\tau_K\,\mathcal{L}u_h\right| = \left|\sum_{K\in\mathcal{T}_h}\int_K \sqrt{\tau_K}f\,\sqrt{\tau_K}\mathcal{L}u_h\right|$$

$$\overset{\text{(Cauchy–Schwarz)}}{\leq} \sum_{K\in\mathcal{T}_h}\|\sqrt{\tau_K}f\|_{L^2(K)}\,\|\sqrt{\tau_K}\mathcal{L}u_h\|_{L^2(K)}$$

$$\overset{\text{(Young)}}{\leq} \sum_{K\in\mathcal{T}_h}\|\sqrt{\tau_K}f\|^2_{L^2(K)} + \frac{1}{4}\,\|\sqrt{\tau_K}\mathcal{L}u_h\|^2_{L^2(K)}$$

To wrap-up, $a_h(u_h, u_h) = F_h(u_h)$ implies:

$$
\begin{aligned}
\|u_h\|^2_{\mathsf{GLS}} &= \int_\Omega \mu |\nabla u_h|^2 + \int_\Omega \gamma\, u_h^2 + \sum_{K\in\mathcal{T}_h} \int_K \tau_K\, (\mathcal{L}u_h)^2 \\
&\le \left[\left\| \frac{1}{\sqrt{\gamma}} f \right\|^2_{L^2(\Omega)} + \sum_{K\in\mathcal{T}_h} \|\sqrt{\tau_K} f\|^2_{L^2(\Omega)} \right] \\
&\quad + \frac{1}{4} \left[\int_\Omega \gamma\, u_h^2 + \sum_{K\in\mathcal{T}_h} \int_\Omega \tau_K\, (\mathcal{L}u_h)^2 \right] \\
&\le \underbrace{\left(\frac{1}{\gamma_0} + \max_{K\in\mathcal{T}_h} \tau_K \right)}_{C \text{ (if } \tau_K \text{ uniformly bounded w.r.t. } h)} \|f\|^2_{L^2(\Omega)} + \frac{1}{4} \|u_h\|^2_{\mathsf{GLS}}
\end{aligned}
$$

$\rightarrow$ $$\boxed{\|u_h\|^2_{\mathsf{GLS}} \le \frac{4}{3} C\, \|f\|^2_{L^2(\Omega)}}$$ **Stability**

On the choice of τ_K (stabilization parameter)

First choice is (6): $\tau_K(\mathbf{x}) = \delta \dfrac{h_K}{|\mathbf{b}(\mathbf{x})|}$, with $\delta > 0$ to be chosen.

Alternative choice:

$$\tau_K(\mathbf{x}) = \frac{h_K}{2|\mathbf{b}(\mathbf{x})|}\xi(\mathbb{P}\mathrm{e}_K) \qquad (10)$$

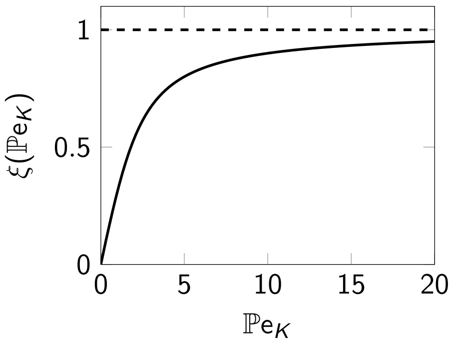
where:

$$\xi(\theta) = \coth(\theta) - \frac{1}{\theta}$$

$$\mathbb{P}\mathrm{e}_K(\mathbf{x}) = \frac{|\mathbf{b}(\mathbf{x})|}{2\,\mu(\mathbf{x})} h_K$$

(local Peclet number)

were $\frac{|\mathbf{b}(\mathbf{x})|}{2\,\mu(\mathbf{x})}$ is the global/physical Peclet number



Remark

Note that since $\lim_{\theta\to+\infty}\xi(\theta)=1$, if $\mathbb{P}\mathrm{e}_K(\mathbf{x})\gg 1$, τ_K defined in (10) reduces, in the limit, to (6) with $\delta=1/2$.

Moreover, if $\theta\to 0$, then $\xi(\theta)=\theta/3+o(\theta)$, therefore when $\mathbb{P}\mathrm{e}_K(\mathbf{x})\ll 1$, we have $\tau_K(\mathbf{x})\to 0$ and **no stabilization** is needed (pure Galerkin works fine!).

Remark

Note also (more tricky) that, in 1D, choice (10) coincides with the famous **Scharfetter-Gummel** stabilization scheme, a second order scheme that is nodally exact (see [Q], Secs. 13.8.7 and 13.6).

[Q] A. Quarteroni, *Numerical Models for Differential Problems*, 3rd Ed. , Springer, 2018

Remark: polynomial of degree $r \geq 1$ arbitrary

(10) modifies as:

$$\mathbb{P}\mathrm{e}_K^r = \frac{|\mathbf{b}(\mathbf{x})|}{2\,\mu(\mathbf{x})\,r} h_K$$

and then:

$$\tau_K(\mathbf{x}) = \frac{h_K}{2\,|\mathbf{b}(\mathbf{x})|\,r} \xi(\mathbb{P}\mathrm{e}_K^r(\mathbf{x}))$$

Convergence of GLS

To state the convergence result for GLS, we need the following inequality, known as *inverse inequality*.

Inverse inequality

$$\sum_{K \in \mathcal{T}_h} h_K^2 \int_K |\Delta v_h|^2 dK \leq C_0 \|\nabla v_h\|^2_{\mathrm{L}^2(\Omega)} \quad \forall v_h \in X_h^r. \tag{11}$$

Theorem (Convergence of GLS)

Assume that the space V_h satisfies the following local approximation property: for each $v \in V \cap H^{r+1}(\Omega)$, there exists a function $\hat{v}_h \in V_h$ s.t.

$$\|v - \hat{v}_h\|_{\mathrm{L}^2(K)} + h_K |v - \hat{v}_h|_{\mathrm{H}^1(K)} + h_K^2 |v - \hat{v}_h|_{\mathrm{H}^2(K)} \leq C h_K^{r+1} |v|_{\mathrm{H}^{r+1}(K)} \quad (12)$$

for each $K \in \mathcal{T}_h$. Moreover, we suppose that for each $K \in \mathcal{T}_h$ the local Péclet number of K satisfies

$$\mathbb{P}e_K(\mathbf{x}) = \frac{|\mathbf{b}(\mathbf{x})|\, h_K}{2\mu} > 1 \quad \forall \mathbf{x} \in K, \quad (13)$$

that is, we are in the pre-asymptotic regime. Finally, we suppose that the inverse inequality holds and that the stabilization parameter satisfies the relation $0 < \delta \leq 2C_0^{-1}$.
Then, as long as $u \in H^{r+1}(\Omega)$, the following super-optimal estimate holds:

$$\|u_h - u\|_{GLS} \leq C h^{r+1/2} |u|_{\mathrm{H}^{r+1}(\Omega)}. \quad (14)$$

Proof.
First of all, we rewrite the error as follows

$$e_h = u_h - u = \sigma_h - \eta, \tag{15}$$

with $\sigma_h = u_h - \hat{u}_h$, $\eta = u - \hat{u}_h$, where $\hat{u}_h \in V_h$ is a function that depends on u and that satisfies property (12). If, for instance, $V_h = X_h^r \cap H_0^1(\Omega)$, we can choose $\hat{u}_h = \Pi_h^r u$, that is the finite element interpolant of u.

We start by estimating the norm $\|\sigma_h\|_{GLS}$. By exploiting the strong consistency of the GLS scheme, we obtain

$$||\sigma_h||^2_{GLS} = a_h(\sigma_h, \sigma_h) = a_h(u_h - u + \eta, \sigma_h) = a_h(\eta, \sigma_h).$$

Now, thanks to the homogeneous Dirichlet boundary conditions it follows that, by adding and subtracting the term $\sum_{K\in\mathcal{T}_h}(\eta,\mathcal{L}\sigma_h)_K$, suitable computations lead to:

$$a_h(\eta,\sigma_h) = \mu\int_\Omega \nabla\eta\cdot\nabla\sigma_h\,d\Omega - \int_\Omega \eta\,\mathbf{b}\cdot\nabla\sigma_h\,d\Omega + \int_\Omega \sigma\,\eta\,\sigma_h\,d\Omega$$

$$+\sum_{K\in\mathcal{T}_h}\delta\Big(\mathcal{L}\eta,\frac{h_K}{|\mathbf{b}|}\mathcal{L}\sigma_h\Big)_{\mathrm{L}^2(K)}$$

$$= \underbrace{\mu(\nabla\eta,\nabla\sigma_h)_{\mathrm{L}^2(\Omega)}}_{(\mathrm{I})} \underbrace{-\sum_{K\in\mathcal{T}_h}(\eta,\mathcal{L}\sigma_h)_{\mathrm{L}^2(K)}}_{(\mathrm{II})} + \underbrace{2(\gamma\,\eta,\sigma_h)_{\mathrm{L}^2(\Omega)}}_{(\mathrm{III})}$$

$$+\underbrace{\sum_{K\in\mathcal{T}_h}(\eta,-\mu\Delta\sigma_h)_{\mathrm{L}^2(K)}}_{(\mathrm{IV})} + \underbrace{\sum_{K\in\mathcal{T}_h}\delta\left(\mathcal{L}\eta,\frac{h_K}{|\mathbf{b}|}\mathcal{L}\sigma_h\right)_{\mathrm{L}^2(K)}}_{(\mathrm{V})}.$$

We now bound the terms (I)-(V) separately.

By carefully using the Cauchy-Schwarz and Young inequalities we obtain

$$|(\mathrm{I})| = |\mu(\nabla\eta, \nabla\sigma_h)_{\mathrm{L}^2(\Omega)}| \leq \frac{\mu}{4}||\nabla\sigma_h||^2_{\mathrm{L}^2(\Omega)} + \mu\|\nabla\eta\|^2_{\mathrm{L}^2(\Omega)},$$

$$\begin{aligned}
|(\mathrm{II})| &= \left|\sum_{K\in\mathcal{T}_h}(\eta, L\sigma_h)_{\mathrm{L}^2(K)}\right| \\
&= \left|\sum_{K\in\mathcal{T}_h}\left(\sqrt{\frac{|\mathbf{b}|}{\delta\, h_K}}\,\eta, \sqrt{\frac{\delta\, h_K}{|\mathbf{b}|}}\,L\sigma_h\right)_{\mathrm{L}^2(K)}\right| \\
&\leq \frac{1}{4}\sum_{K\in\mathcal{T}_h}\delta\left(\frac{h_K}{|\mathbf{b}|}L\sigma_h, L\sigma_h\right)_{\mathrm{L}^2(K)} + \sum_{K\in\mathcal{T}_h}\left(\frac{|\mathbf{b}|}{\delta\, h_K}\eta, \eta\right)_{\mathrm{L}^2(K)},
\end{aligned}$$

$$\begin{aligned}
|(\mathrm{III})| &= 2|(\gamma\,\eta, \sigma_h)_{\mathrm{L}^2(\Omega)}| = 2|(\sqrt{\gamma}\,\eta, \sqrt{\gamma}\,\sigma_h)_{\mathrm{L}^2(\Omega)}| \\
&\leq \frac{1}{2}\,\|\sqrt{\gamma}\,\sigma_h\|^2_{\mathrm{L}^2(\Omega)} + 2\,\|\sqrt{\gamma}\,\eta\|^2_{\mathrm{L}^2(\Omega)}.
\end{aligned}$$

For the term (IV), thanks again to the Cauchy-Schwarz and Young inequalities, hypothesis (13) and the inverse inequality (11), we obtain

$$
\begin{aligned}
|(\mathrm{IV})| &= \left| \sum_{K\in\mathcal{T}_h} (\eta, -\mu\Delta\sigma_h)_{\mathrm{L}^2(K)} \right| \\
&\leq \frac{1}{4} \sum_{K\in\mathcal{T}_h} \delta\,\mu^2 \Big(\frac{h_K}{|\mathbf{b}|}\,\Delta\sigma_h, \Delta\sigma_h\Big)_{\mathrm{L}^2(K)} \\
&+ \sum_{K\in\mathcal{T}_h} \Big(\frac{|\mathbf{b}|}{\delta\, h_K}\eta, \eta\Big)_{\mathrm{L}^2(K)} \\
&\leq \frac{1}{8}\,\delta\,\mu \sum_{K\in\mathcal{T}_h} h_K^2\ (\Delta\sigma_h, \Delta\sigma_h)_{\mathrm{L}^2(K)} + \sum_{K\in\mathcal{T}_h} \Big(\frac{|\mathbf{b}|}{\delta\, h_K}\eta, \eta\Big)_{\mathrm{L}^2(K)} \\
&\leq \frac{\delta\, C_0\, \mu}{8} \|\nabla\sigma_h\|^2_{\mathrm{L}^2(\Omega)} + \sum_{K\in\mathcal{T}_h} \Big(\frac{|\mathbf{b}|}{\delta\, h_K}\eta, \eta\Big)_{\mathrm{L}^2(K)}.
\end{aligned}
$$

Term (V) can finally be bounded once again thanks to the Cauchy-Schwarz and Young inequalities as follows

$$
\begin{aligned}
|(\mathrm{V})| &= \left| \sum_{K\in\mathcal{T}_h} \delta \Big(L\eta, \frac{h_K}{|\mathbf{b}|} L\sigma_h \Big)_{\mathrm{L}^2(K)} \right| \\
&\leq \frac{1}{4} \sum_{K\in\mathcal{T}_h} \delta \left(\frac{h_K}{|\mathbf{b}|} L\sigma_h, L\sigma_h \right)_{\mathrm{L}^2(K)} + \sum_{K\in\mathcal{T}_h} \delta \left(\frac{h_K}{|\mathbf{b}|} L\eta, L\eta \right)_{\mathrm{L}^2(K)}.
\end{aligned}
$$

Thanks to these upper bounds, we obtain the following estimate

$$
\begin{aligned}
&\|\sigma_h\|^2_{GLS} = a_h(\eta, \sigma_h) \leq \frac{1}{4}\, \|\sigma_h\|^2_{GLS} \\
&+ \frac{1}{4}\left(\|\sqrt{\gamma}\, \sigma_h\|^2_{\mathrm{L}^2(\Omega)} + \sum_{K \in \mathcal{T}_h} \delta \left(\frac{h_K}{|\mathbf{b}|}\, L\sigma_h, L\sigma_h \right)_{\mathrm{L}^2(K)} \right) + \frac{\delta\, C_0\, \mu}{8}\, \|\nabla \sigma_h\|^2_{\mathrm{L}^2(\Omega)} \\
&+ \underbrace{\mu\, \|\nabla \eta\|^2_{\mathrm{L}^2(\Omega)} + 2 \sum_{K \in \mathcal{T}_h} \left(\frac{|\mathbf{b}|}{\delta\, h_K} \eta, \eta \right)_{\mathrm{L}^2(K)} + 2\, \|\sqrt{\gamma}\, \eta\|^2_{\mathrm{L}^2(\Omega)} + \sum_{K \in \mathcal{T}_h} \delta \left(\frac{h_K}{|\mathbf{b}|}\, L\eta, L\eta \right)_{\mathrm{L}^2(K)}}_{\mathcal{E}(\eta)} \\
&\leq \frac{1}{2}\, \|\sigma_h\|^2_{GLS} + \mathcal{E}(\eta),
\end{aligned}
$$

having exploited, in the last passage, the assumption that $\delta \leq 2 C_0^{-1}$.

Then, we can state that

$$
\|\sigma_h\|^2_{GLS} \leq 2\, \mathcal{E}(\eta).
$$

We now estimate the term $\mathcal{E}(\eta)$, by bounding each of its summands separately. To this end, we will basically use the local approximation property (12) and the requirement formulated in (13) on the local Péclet number $\mathbb{P}e_K$.

Moreover, we observe that the constants C, introduced in the remainder, depend neither on h nor on $\mathbb{P}e_K$, but can depend on other quantities such as the constant γ_1 in (9), the reaction constant σ, the norm $||\mathbf{b}||_{L^\infty(\Omega)}$, the stabilization parameter δ.

We then have

$$
\begin{aligned}
\mu \, \|\nabla \eta\|^2_{\mathrm{L}^2(\Omega)} &\leq C \, \mu \, h^{2r} \, |u|^2_{\mathrm{H}^{r+1}(\Omega)} \\
&\leq C \, \frac{||\mathbf{b}||_{L^\infty(\Omega)} \, h}{2} \, h^{2r} \, |u|^2_{\mathrm{H}^{r+1}(\Omega)} \leq C \, h^{2r+1} \, |u|^2_{\mathrm{H}^{r+1}(\Omega)}
\end{aligned}
$$

$$2 \sum_{K\in\mathcal{T}_h} \left(\frac{|\mathbf{b}|}{\delta\, h_K}\eta, \eta \right)_{\mathrm{L}^2(K)} \leq C\, \frac{||\mathbf{b}||_{L^\infty(\Omega)}}{\delta} \sum_{K\in\mathcal{T}_h} \frac{1}{h_K}\, h_K^{2(r+1)}\, |u|^2_{\mathrm{H}^{r+1}(K)}$$

$$\leq C\, h^{2r+1}\, |u|^2_{\mathrm{H}^{r+1}(\Omega)},$$

$$2\, \|\sqrt{\gamma}\,\eta\|^2_{\mathrm{L}^2(\Omega)} \leq 2\,\gamma_1\, \|\eta\|^2_{\mathrm{L}^2(\Omega)} \leq C\, h^{2(r+1)}\, |u|^2_{\mathrm{H}^{r+1}(\Omega)},$$

having exploited, for controlling the third summand, the assumption (9).

Finding an upper bound for the fourth summand of $\mathcal{E}(\eta)$ is slightly more difficult: first, by elaborating on the term $L\eta$, we have

$$
\begin{aligned}
&\sum_{K\in\mathcal{T}_h} \delta \left(\frac{h_K}{|\mathbf{b}|}\, L\eta, L\eta \right)_{\mathrm{L}^2(K)} \\
&= \sum_{K\in\mathcal{T}_h} \delta \left\| \sqrt{\frac{h_K}{|\mathbf{b}|}}\, L\eta \right\|^2_{\mathrm{L}^2(K)} \\
&= \sum_{K\in\mathcal{T}_h} \delta \left\| -\mu \sqrt{\frac{h_K}{|\mathbf{b}|}}\, \Delta\eta + \sqrt{\frac{h_K}{|\mathbf{b}|}}\, \mathrm{div}(\mathbf{b}\eta) + \sigma \sqrt{\frac{h_K}{|\mathbf{b}|}}\, \eta \right\|^2_{\mathrm{L}^2(K)} \qquad (16) \\
&\leq C \sum_{K\in\mathcal{T}_h} \delta \left(\left\| \mu \sqrt{\frac{h_K}{|\mathbf{b}|}}\, \Delta\eta \right\|^2_{\mathrm{L}^2(K)} + \left\| \sqrt{\frac{h_K}{|\mathbf{b}|}}\, \mathrm{div}(\mathbf{b}\eta) \right\|^2_{\mathrm{L}^2(K)} \right. \\
&\left. + \left\| \sigma \sqrt{\frac{h_K}{|\mathbf{b}|}}\, \eta \right\|^2_{\mathrm{L}^2(K)} \right).
\end{aligned}
$$

Now, with a similar computation to the one performed to obtain estimates (28) and (29), it is easy to prove that the second and third summands of the left-hand side of (16) can be bounded using a term of the form $C\,h^{2r+1}\,|u|^2_{\mathrm{H}^{r+1}(\Omega)}$, for a suitable choice of the constant C. For the first summand, we have

$$\sum_{K\in\mathcal{T}_h}\delta\left\|\mu\sqrt{\frac{h_K}{|\mathbf{b}|}}\,\Delta\eta\right\|^2_{\mathrm{L}^2(K)} \leq \sum_{K\in\mathcal{T}_h}\delta\,\frac{h_K^2\,\mu}{2}\,\|\Delta\eta\|^2_{\mathrm{L}^2(K)}$$
$$\leq C\,\delta\,||\mathbf{b}||_{L^\infty(\Omega)}\sum_{K\in\mathcal{T}_h}h_K^3\,\|\Delta\eta\|^2_{\mathrm{L}^2(K)} \leq C\,h^{2r+1}\,|u|^2_{\mathrm{H}^{r+1}(\Omega)},$$

having again used conditions (12) and (13). The latter bound allows us to conclude that

$$\mathcal{E}(\eta) \leq C\,h^{2r+1}\,|u|^2_{\mathrm{H}^{r+1}(\Omega)},$$

that is

$$\|\sigma_h\|_{GLS} \leq C\,h^{r+1/2}\,|u|_{\mathrm{H}^{r+1}(\Omega)}. \tag{17}$$

Reverting to (15), to obtain the desired estimate for the norm $\|u_h - u\|_{GLS}$ we still have to estimate $\|\eta\|_{GLS}$. This evidently leads to estimating three contributions as in (28), (29) and (16), and eventually produces

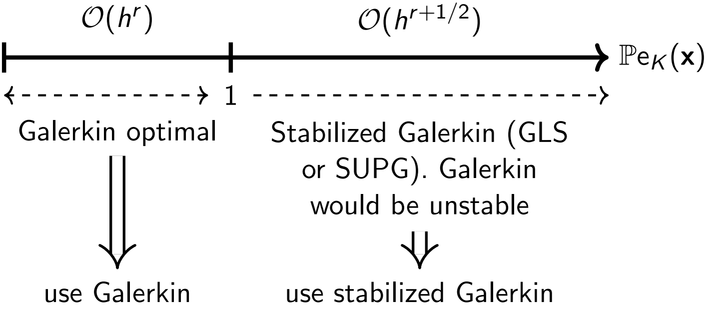
$$\|\eta\|_{GLS} \leq C\, h^{r+1/2}\, |u|_{\mathrm{H}^{r+1}(\Omega)}.$$

The desired estimate (14) follows by combining this result with (17). $\square$

Rationale for (13): use GLS only beyond the asymptotic regime.
In the asymptotic regime, i.e. if $\mathbb{P}e_K < 1$, use standard Galerkin.

Notice that (14) is **"super-optimal"**: $h^{r+1/2}$ instead of h^r as in Galerkin.
The reason is (13) (a lower limit for h).

We are not in the asymptotic regime for $h \to 0$ ($\implies \mathbb{P}e_K \to 0$).



[Q] A. Quarteroni, *Numerical Models for Differential Problems*, 3rd Ed. , Springer, 2018

DG methods for diffusion-transport equations (not for exam)

The Discontinuous Galerkin method can be extended to the diffusion-transport-reaction problem in conservation form:

$$\begin{cases} -\operatorname{div}(\mu \nabla u + \mathbf{b} u) + \sigma u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \tag{18}$$

We introduce the following space:

$$W_\delta^0 = \left\{ v_\delta \in L^2(\Omega) : \ v_\delta|_K \in H^1(K) \, \forall K \in \mathcal{T}_h, \ v_\delta|_{\partial\Omega} = 0 \right\}$$

The Discontinuous Galerkin formulation reads as follow:

find $u_\delta \in W_\delta^0$ s.t.

$$\sum_{K\in\mathcal{T}_h} (\mu \nabla u_\delta, \nabla v_\delta)_{L^2(K)} - \sum_{e\in\mathcal{E}_\delta} \int_e [\![v_\delta]\!] \cdot \{\!\{\mu\nabla u_\delta\}\!\} - \theta \sum_{e\in\mathcal{E}_\delta} \int_e [\![u_\delta]\!] \, \{\!\{\mu\nabla v_\delta\}\!\}$$
$$+ \sum_{e\in\mathcal{E}_\delta} \int_e \overline{\gamma}\, [\![u_\delta]\!] \cdot [\![v_\delta]\!] - \sum_{K\in\mathcal{T}_h} (\mathbf{b} u_\delta, \nabla v_\delta)_{L^2(K)} + \sum_{e\in\mathcal{E}_\delta} \int_e \{\!\{\mathbf{b} u_\delta\}\!\}_{\mathbf{b}} \cdot [\![v_\delta]\!]$$
$$+ \sum_{K\in\mathcal{T}_h} (\sigma u_\delta, v_\delta)_{L^2(K)} = \sum_{K\in\mathcal{T}_h} (f, v_\delta)_{L^2(K)} \quad , \tag{19}$$

where $\mathcal{E}_\delta \equiv \mathcal{F}_h$ is the set of the edges of the elements $\{K\}$, $\overline{\gamma}$ is the DG stabilization function (see Lecture 03, slide 12) and where

$$\{\!\{\mathbf{b} u_\delta\}\!\}_{\mathbf{b}} = \begin{cases} \mathbf{b} u_\delta^+ & \text{if} \quad \mathbf{b}\cdot\mathbf{n}^+ > 0 \\ \mathbf{b} u_\delta^- & \text{if} \quad \mathbf{b}\cdot\mathbf{n}^+ < 0 \\ \mathbf{b}\{\!\{u_\delta\}\!\} & \text{if} \quad \mathbf{b}\cdot\mathbf{n}^+ = 0 \, . \end{cases} \tag{20}$$

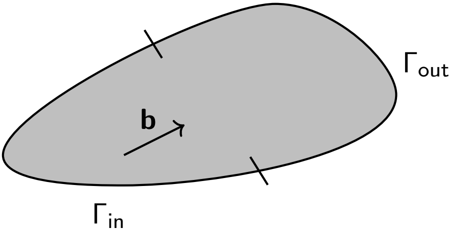
Observe that $\{\!\{\mathbf{b} u_\delta\}\!\}_{\mathbf{b}} \cdot [\![v_\delta]\!] = 0$ if $\mathbf{b}\cdot\mathbf{n}^+ = 0$.

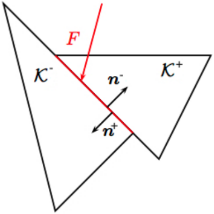
If the diffusion-transport-reaction equation is written in non-conservative form, by $\mathbf{b} \cdot \nabla u = \mathrm{div}(\mathbf{b}u) - \mathrm{div}(\mathbf{b})u$ it is sufficient to modify (19) by substituting the term

$$\sum_{m=1}^{M} (\sigma u_\delta, v_\delta)_{\Omega_m} \quad \text{with} \quad \sum_{m=1}^{M} (\eta u_\delta, v_\delta)_{\Omega_m},$$

where $\eta(\mathbf{x}) = \sigma(\mathbf{x}) - \mathrm{div}(\mathbf{b}(\mathbf{x}))$.

This time we suppose that there exists a positive constant $\eta_0 > 0$ so that $\eta(\mathbf{x}) \geq \eta_0$ for almost every $\mathbf{x} \in \Omega$.





Some numerical tests

We now present some numerical solutions obtained using linear finite elements for the following two-dimensional diffusion-transport problem

$$\begin{cases} -\mu\Delta u + \mathbf{b}\cdot\nabla u = f \quad \text{in } \Omega = (0,1)\times(0,1), \\ u = g \quad \text{on } \partial\Omega, \end{cases} \tag{21}$$

where $\mathbf{b} = (-1,1)^T$. To start with let us consider the following constant data: $f \equiv 1$ and $g \equiv 0$. In this case the solution is characterized by a boundary layer near the edges $x = 0$ and $y = 1$.

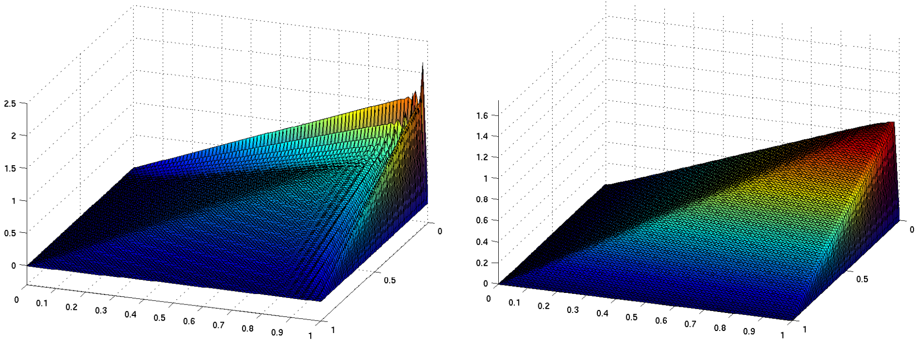


Figure: Approximation of problem (21) with $\mu = 10^{-3}$, $h = 1/80$, using the standard (left) and GLS (right) Galerkin method. The corresponding local Péclet number is $\mathbb{P}\mathrm{e}_K = 8.84$

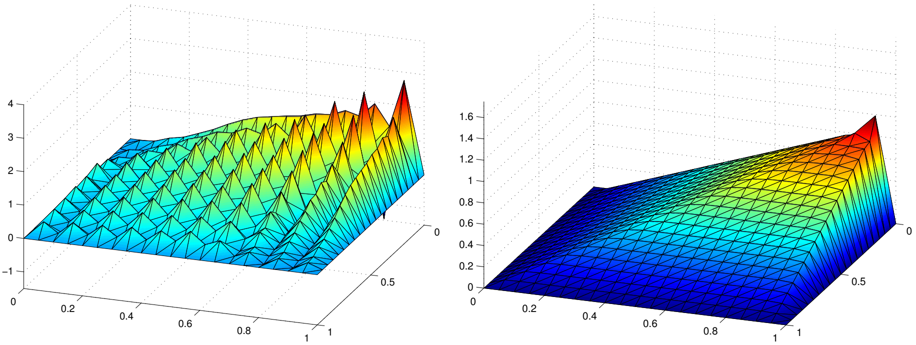
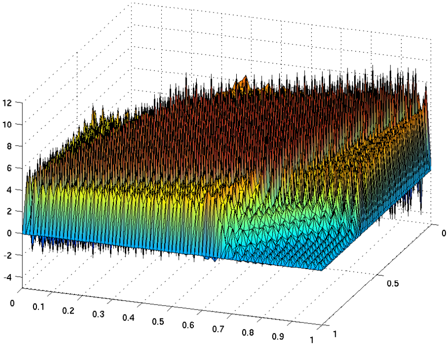


Figure: Approximation of problem (21) with $\mu = 10^{-3}$, $h = 1/20$, using the standard (left) and GLS (right) Galerkin method. The corresponding local Péclet number is $\mathbb{P}\mathrm{e}_K = 35.35$



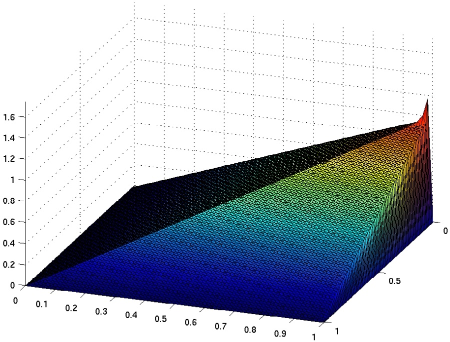


Figure: Approximation of problem (21) with $\mu = 10^{-5}$, $h = 1/80$, using the standard (left) and GLS (right) Galerkin method. The corresponding local Péclet number is $\mathbb{P}\mathrm{e}_K = 883.88$

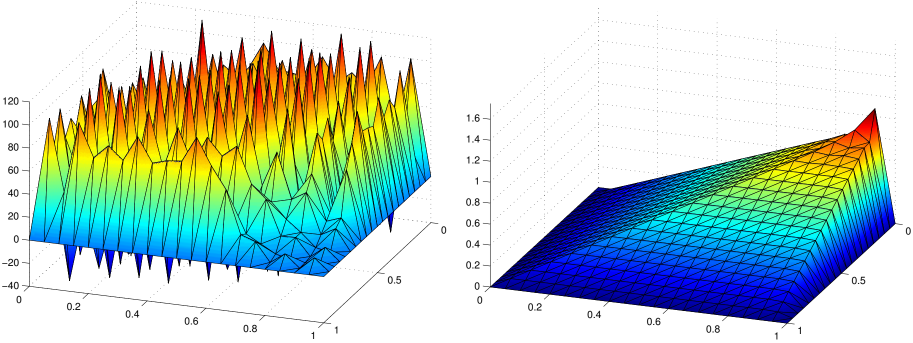


Figure: Approximation of problem (21) with $\mu = 10^{-5}$, $h = 1/20$, using the standard (left) and GLS (right) Galerkin method. The corresponding local Péclet number is $\mathbb{P}\mathrm{e}_K = 3535.5$

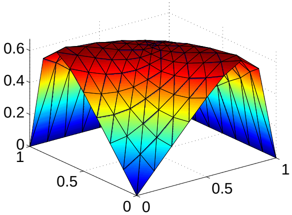
Let us now set $\mathbf{b} = (1,1)^T$ and choose forcing term f and the boundary data g in such a way that

$$u(x,y) = x + y(1-x) + \frac{\mathrm{e}^{-1/\mu} - \mathrm{e}^{-(1-x)(1-y)/\mu}}{1 - \mathrm{e}^{-1/\mu}}$$

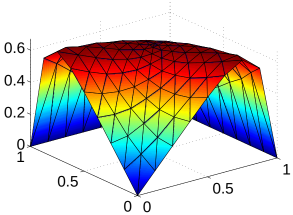
is the exact solution.

The corresponding Péclet number is $\mathbb{P}e = (\sqrt{2}\mu)^{-1}$.

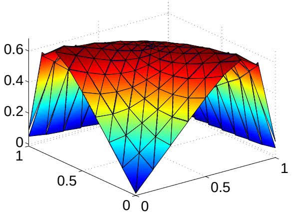
For small values of the viscosity μ, this solution features a boundary layer near the edges $x = 1$ and $y = 1$.



(a) Standard Galerkin method

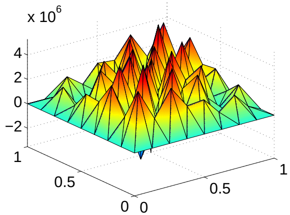


(b) SUPG method

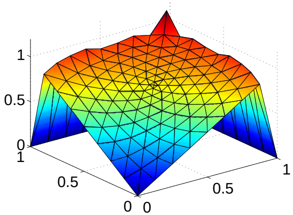


(c) DG method

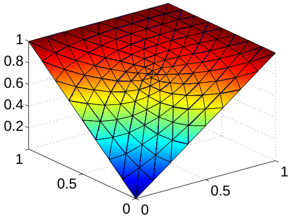
Figure: The approximate solution of problem (21) for $\mu = 10^{-1}$. Triangular grid with $h \approx 1/8$ and piecewise linear finite elements ($r = 1$).



(a) Standard Galerkin method

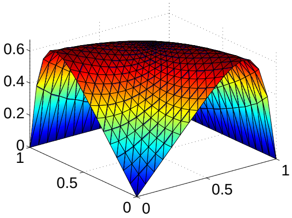


(b) SUPG method

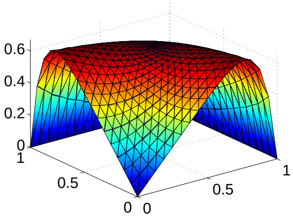


(c) DG method

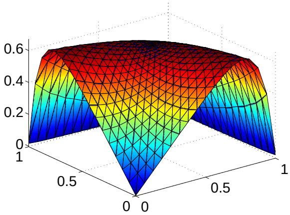
Figure: The approximate solution of problem (21) for $\mu = 10^{-9}$. Triangular grid with $h \approx 1/8$ and piecewise linear finite elements ($r = 1$).



(a) Standard Galerkin method

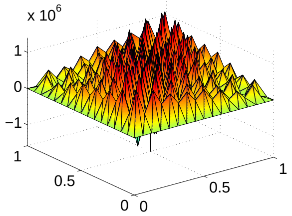


(b) SUPG method

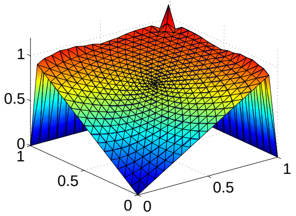


(c) DG method

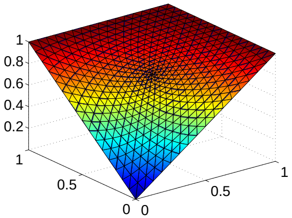
Figure: The approximate solution of problem (21) for $\mu = 10^{-1}$. Triangular grid with $h \approx 1/16$ and piecewise linear finite elements ($r = 1$).



(a) Standard Galerkin method



(b) SUPG method



(c) DG method

Figure: The approximate solution of problem (21) for $\mu = 10^{-9}$. Triangular grid with $h \approx 1/16$ and piecewise linear finite elements ($r = 1$).

Finally, we consider a pure transport problem, that is $\mathbf{b} \cdot \nabla u = f$ in $\Omega = (0,1)^2$ with $u = g$ su Γ^-, $\mathbf{b} = (1,1)$, with f and g chosen in such a way that the exact solution is $u(x,y) = 1 + \sin(\pi(x+1)(y+1)^2/8)$.

We solve this problem by the DG method with piecewise polynomials of degree $r = 1, 2, 3$ and 4 on a sequence of uniform triangular grids with gridsize h.

The DG method provides the following error estimate [1]

$$\begin{aligned} |||u - u_h||| &= \left(\|u - u_h\|^2_{L^2(\Omega)} + \sum_{e \in \mathcal{E}_h} \|s_e^{1/2}[u - u_h]\|^2_{0,e} \right)^{1/2} \\ &\leq C h^{r+1/2} \|u\|_{H^{r+1}(\Omega)}, \end{aligned} \tag{22}$$

where $s_e = \alpha |\mathbf{b} \cdot \mathbf{n_e}|$ is a suitable stabilization term, where α is a positive constant independent of h and e. $\mathcal{E}_h$ is the set of all the edges of the triangulation and C is a positive constant.

[1] see e.g. F. Brezzi, L. D. Marini and E. Süli, “Discontinuos Galerkin methods for first-order hyperbolic problems”, *Math. Models Methods Appl. Sci.* (2004)

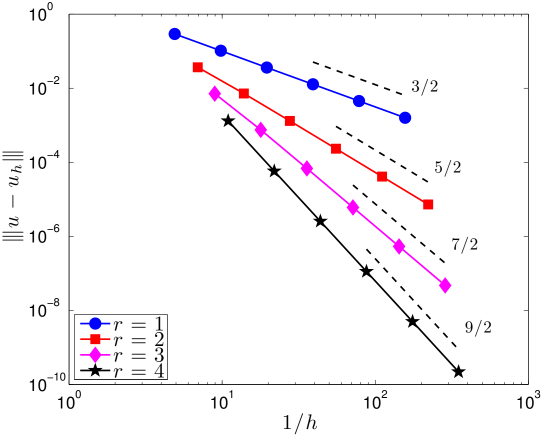


Figure: Approximation error (in the energy norm (22)) vs number of degrees of freedom for finite elements of degree $r = 1, 2, 3, 4$

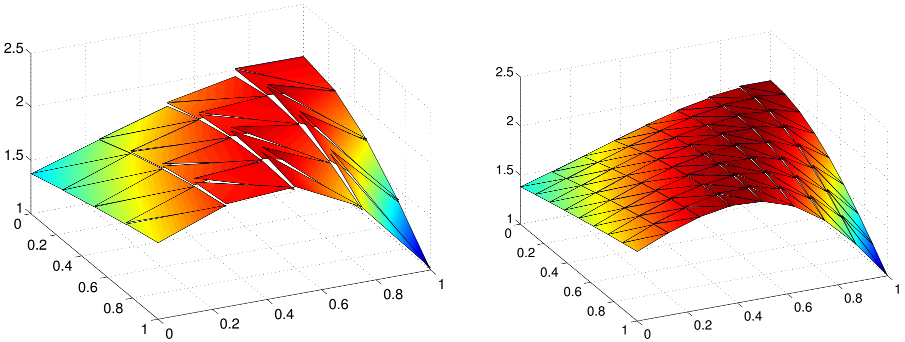


Figure: Finite element solutions obtained on a uniform grid with gridsize $h = 1/4$ (left) and $h = 1/8$ (right)