Finite element methods

Yuji Nakatsukasa

Many slides by Ricardo Ruiz Baier
Some pictures from Andy Wathen
Mathematical Institute, Oxford

Computational Techniques
InFoMM Centre for Doctoral Training
Michaelmas Term 2019, Week 6

November 21, 2019

University of Oxford
Mathematical Institute

Oxford Mathematics

Finite element methods: references

- References (among many):
  - H. Elman, D. Silvester, A. Wathen, Finite Elements and Fast Iterative Solvers, OUP, 2014
  - P. E. Farrell, Finite Element Methods for PDEs, C6.4 lecture notes
  - E. Süli, Lecture Notes for FEM for PDEs `people.maths.ox.ac.uk/suli/fem.pdf`
  - Brenner and Scott, The Mathematical Theory of Finite Element Methods, Springer, 2007 (advanced, mathematical analysis)
- For software, we suggest FEniCS (demo session later by Fede Danieli) or IFISS (Elman-Silvester-Wathen book)
- For advanced questions, we suggest asking our amazing local experts!
  - P. E. Farrell (theory, programming & applications)
  - E. Süli (analysis, theory)
  - A. Wathen (preconditioning, LA aspects)

Classification of PDEs

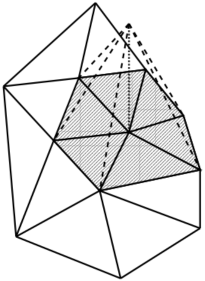
$$au_{xx} + bu_{xy} + cu_{yy} + du_x + eu_y + fu = g$$

- elliptic: $b^2 - 4ac < 0$
  e.g. Poisson problem $\nabla^2 u = f$
- parabolic: $b^2 - 4ac = 0$
  e.g. heat equation $u_t = \nabla^2 u - f$
- hyperbolic $b^2 - 4ac > 0$
  e.g. wave equation $u_{tt} = c^2 u_{xx}$

Finite elements: essentials

e.g. consider Poisson problem $\nabla^2 u = f$

- Approximate solution u with piecewise polynomial $\hat{u}(x) = \sum_{i=1}^{n} c_i \phi_i(x)$
  e.g. $\phi_i(x)$: hat function
- Integration by parts+divergence thm to 'move' one derivative, 'relax' smoothness requirement
- Find c_i via 'weak solution' by requiring Galerkin condition: 'residual is orthogonal to test functions', in LA terms, $Q^T(Ax - b) = 0$ (recall least-squares $\min_x \|Ax - b\|$, and CG $Q^T(AQy - b) = 0$)



Poisson problem in weak form I

First, we stick to the Poisson problem on a bounded domain (in strong form)

$$-\nabla^2 u = f \quad \text{in } \Omega \subset \mathbb{R}^d, \qquad u = 0 \text{ on } \partial\Omega.$$

Weak formulation

- multiply by a test function $v \in V$
- integrate by parts

  $\int_\Omega \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} (\nabla u \cdot \mathbf{n}) v \, ds = \int_\Omega f v \, dx$
- BCs $\to$ $V = \{w \in H^1(\Omega) : w = 0 \text{ on } \partial\Omega\}$
  (roughly, $H^k(\Omega)$: k-times differentiable with kth der$\in L^2(\Omega)$)
- find $u \in V$ such that

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx, \qquad \forall v \in V$$

Weak form for poisson, step by step I

$$-\nabla^2 u = f \quad \text{in } \Omega \subset \mathbb{R}^d, \qquad u = 0 \text{ on } \partial\Omega.$$

1. Multiply by a test function $v \in V$ and integrate:

$$-\int_\Omega v\nabla^2 u dx = \int_\Omega fv dx$$

2. Integrate by parts: first recall product rule

$$v\nabla^2 u + \nabla v \cdot \nabla u = \nabla \cdot (v\nabla u)$$

Integrate $-\int_\Omega v\nabla^2 u dx = \int_\Omega \nabla v \cdot \nabla u dx - \int_\Omega \nabla \cdot (v\nabla u) dx$. By diver. thm. $\int_\Omega \nabla \cdot (v\nabla u) dx = \int_{\partial\Omega} (\nabla u \cdot \mathbf{n}) v \, ds$ ($\mathbf{n}$: outward normal vec)

$$\int_\Omega \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} (\nabla u \cdot \mathbf{n}) v \, ds = \int_\Omega fv \, dx \tag{1}$$

$\rightarrow$ reduced regularity: before ($u \in C^2(\bar{\Omega})$), after ($u \in C^1(\bar{\Omega})$)

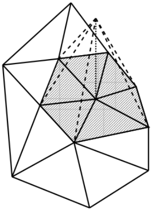
Weak form for poisson, step by step II

3. Take test functions $v = \xi_1(x), \ldots, \xi_n(x)$ in (1) (simplest case: $\phi_i = \xi_i$=hat func) to find $\hat{u}(x) = \sum_{i=1}^{n} c_i \phi_i(x)$ via $n \times n$ linear system $A\mathbf{c} = \mathbf{f}$, where

$$A_{ij} = \int_\Omega \nabla\phi_j \cdot \nabla\xi_i \, dx - \int_{\partial\Omega} (\nabla\phi_j \cdot \mathbf{n})\xi_i \, ds, \quad f_i = \int_\Omega f \, \xi_i \, dx$$

- When $\phi_i = \xi_i$=hat functions, reduced regularity significant–why?
- We'll take $\xi_i = 0$ on boundary $\partial\Omega$

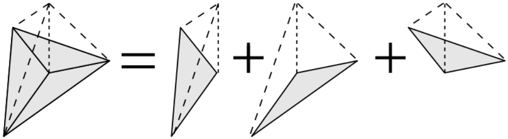
Linear system is sparse and positive definite



Recall $A_{ij} = \int_\Omega \nabla\phi_j \cdot \nabla\xi_j \, dx$

If support of ϕ_j, ξ_j do not overlap, $A_{ij} = 0$
$\Rightarrow A$ highly sparse! exploit in solving $A\mathbf{c} = \mathbf{f}$

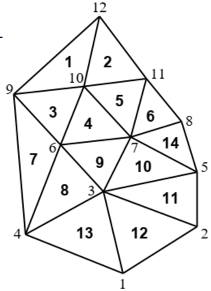
For nonzero entries, compute A_{ij} via splitting into



Sparsity

UNIVERSITY OF OXFORD
Mathematical Institute

What is the sparsity structure of A here?



- Taking $\phi_i = \xi_i = 0$ on $\partial\Omega$, $A_{ij} = \int_\Omega \nabla\phi_j \cdot \nabla\xi_j \, dx$
- Then A symmetric positive definite:

$$v^T A v = \sum_{j=1}^{n} \sum_{i=1}^{n} v_j A_{ji} v_i = \sum_{j=1}^{n} \sum_{i=1}^{n} v_j \Big(\int_\Omega \nabla\phi_j \cdot \nabla\phi_j \, dx \Big) v_i$$

$$= \int_\Omega \left(\sum_{j=1}^{n} v_j \nabla\phi_j \right) \cdot \left(\sum_{i=1}^{n} v_i \nabla\phi_j \right) dx \geq 0.$$

- (preconditioned) conjugate gradient applicable/effective for $A\mathbf{c} = \mathbf{f}$

More generally

- In this special case, $V = H_0^1(\Omega)$
- If $u = g$ on $\partial\Omega$ we rewrite the weak problem: find $u \in V_g$ such that

$$a(u, v) = F(v), \qquad \forall v \in V_0$$

- Trial space: $V_g = \{w \in H^1(\Omega) : w = g \text{ on } \partial\Omega\}$, test space: $V_0 = H_0^1(\Omega)$
- Alternatively: lifting strategy (solve the homogeneous weak form for $u - u_g$ where u_g is a function st. $u_g = g$ on $\partial\Omega$)
- common notation: $(u, v) := \int_\Omega uv\, dx$, $\|v\|_{0,\Omega}^2 = (v, v)$
- u solution of the weak formulation need not belong to $C^2(\bar{\Omega})$, but if it does, then it is a *strong solution*

Theorem

(Lax-Milgram) Let $(V, \|\cdot\|_V)$ be a Hilbert space and V_0 a closed subspace and consider the problem: find $u \in V$ st

$$a(u, v) = F(v), \qquad \forall v \in V_0.$$

Assume

- *$a(\cdot,\cdot)$ is bounded:* $|a(v,w)| \leq C_1 \|v\|_V \|w\|_V$, $v, w \in V$
- *$a(\cdot,\cdot)$ is $V-$elliptic (or coercive):* $a(v,v) \geq C_2 \|v\|_V^2$, $v \in V$
- *$F(\cdot)$ is bounded:* $|F(v)| \leq C_3 \|v\|_V$, $v \in V$

Then the problem is uniquely solvable and $\|u\|_V \leq C_2^{-1} \|F\|_{V'}$.

(But this is not a solution *method*!)

Back to Mr. Poisson

Well-posedness

Let's check it: find $u \in H^1(\Omega)$ such that

$$u = 0 \quad \text{on } \partial\Omega, \qquad \text{and } a(u, v) = F(v) \quad \forall v \in H^1_0(\Omega).$$

- $H^1(\Omega)$ with the norm $\|v\|^2_{1,\Omega} := \|v\|^2_{0,\Omega} + \|\nabla v\|^2_{0,\Omega}$ is a Hilbert space
- the bilinear form is bounded (C-S and norm def.)
- the linear functional is bounded (C-S and norm def.)
- the bilinear form is $H^1(\Omega)-$elliptic (established using Poincaré ineq.)

Cauchy-Schwarz inequality: $|(v, w)| \leq \|v\| \, \|w\|$, for $v, w \in V$

Poincaré inequality: $\|v\|_{0,\Omega} \leq C \, \|\nabla v\|_{0,\Omega}$, for $v \in H^1_0(\Omega)$

Galerkin method

Let's now consider V_h subspace of V, with $\dim V_h = n < \infty$

- Replace V by V_h in the weak form. We get: find $u_h \in V_h$ (an approximation of u) st.

$$a(u_h, v_h) = F(v_h) \qquad \forall v_h \in V_h.$$

- Done. This was Galerkin's method
- It can be reduced to a set of n linear eqns. and n unknowns
- Comparing the "continuous" and "discrete" problems gives the

  Galerkin orthogonality ("strong" consistency)

$$a(u - u_h, v_h) = 0 \qquad \forall v_h \in V_h$$

(using that a, F are unchanged and that a is linear)

Galerkin method

Further properties

$V_h \subset V \Rightarrow$ Lax-Milgram also applicable for the Galerkin problem $\Rightarrow$

- The solution of the Galerkin problem exists and is unique
- The method is uniformly stable wrt h since $\|u_h\|_V \leq C_2^{-1} \|F\|_{V'}$

Céa's estimate: $a(\cdot,\cdot)$ bilinear, continuous and $V-$elliptic. Then

$$\|u - u_h\|_V \leq C_1 C_2^{-1} \inf_{v_h \in V_h} \|u - v_h\|_V$$

Convergence:

$$\lim_{h \to 0} \|u_h - u\|_V = 0,$$

valid if V_h is chosen adequately

Galerkin method

optimality in energy norm

Theorem: $\|\nabla u - \nabla u_h\| = \min\{\|\nabla u - \nabla v_h\| : v_h = \sum_{i=1}^{n} c_i \phi_i\}$,

where $\|\nabla u\|^2 := \int_\Omega (\nabla u \cdot \nabla u) dx (=: a(u,u))$, energy norm

Proof:

$$
\begin{aligned}
\|\nabla u - \nabla u_h\|^2 &= a(u - u_h, u - u_h) = a(u - u_h, u - v_h + v_h - u_h) \\
&= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\
&= a(u - u_h, u - v_h)
\end{aligned}
$$

due to Galerkin orthogonality, since $a(u - u_h, v_h - u_h) = \int_\Omega (\nabla(u - u_h) \cdot \nabla(v_h - u_h)) dx = (r, v_h - u_h)$. By Cauchy-Schwarz,

$$a(u - u_h, u - v_h) \leq \|\nabla(u - u_h)\| \cdot \|\nabla(u - v_h)\|$$

Galerkin method I

Bases and the FEM

- Let $\{\phi_j\}$ be a basis of V_h
- $\Rightarrow$ we have only to guarantee that the Galerkin problem holds for all functions of the basis

$$a(u_h, \phi_i) = F(\phi_i), \qquad i = 1, \ldots, n.$$

- Since $u_h \in V_h$, then $u_h(x) = \sum_{j=1}^{n} c_j \phi_j(x), \qquad$ (with unknown coeffs)
- Then $\sum_{j=1}^{n} c_j a(\phi_j, \phi_i) = F(\phi_i), \qquad i = 1, \ldots, n$
- A: stiffness matrix ($a_{ij} = a(\phi_j, \phi_i)$), $\mathbf{f}$: load vector $f_i = F(\phi_i)$
- $A\mathbf{c} = \mathbf{f}$. If associated to a coercive problem, then A is positive definite

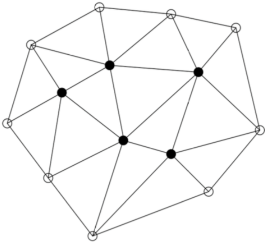
But V_h still not revealed! (which will actually dictate the form of A)

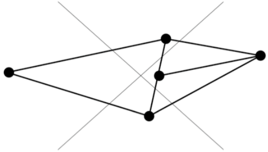
Galerkin method II

Bases and the FEM

Let's "discretize" the remainder of the problem (spaces, weak form, domain)

- Polygonal domain $\Omega \subset \mathbb{R}^2$, partition it into triangles
- If two triangles have some intersection, it is either on common vertex or a common full edge. In particular, two different triangles do not overlap
- h: length of the longest edge of all K in the "regular mesh" $\mathscr{T}_h$





Galerkin method III

Bases and the FEM

- $\mathbb{P}_r$: polynomials of degree r or less. E.g.
  $\mathbb{P}_1 = \{g(\boldsymbol{x}) = a + bx_1 + cx_2, \text{ with } a, b, c \in \mathbb{R}\}$
- $\dim \mathbb{P}_r = (r+1)(r+2)/2$
- On each $K \in \mathscr{T}_h$, v_h is well-defined knowing its value in $\dim \mathbb{P}_r$ points

Finite element space

$$X_h^r = \{v_h \in C^0(\bar{\Omega}) : v_h|_K \in \mathbb{P}_r, \ \forall K \in \mathscr{T}_h\}$$

and the one accounting for the BC

$$\overset{\circ}{X}{}_h^r = \{v_h \in X_h^r : v_h|_{\partial\Omega} = 0\}$$

Lemma

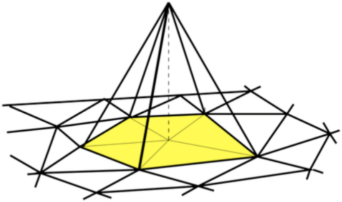
If $v \in C^0(\bar{\Omega})$ and $v \in H^1(K)$ for all $K \in \mathscr{T}_h$, then $v \in H^1(\Omega)$.

For our Poisson problem (with the given BC) we set $V_h = \mathring{X}_h^r$

OK. V_h more or less clear, but what about $\{\phi_j\}$?

Since (in this particular case) $V_h = \mathring{X}_h^r$, each v_h is characterized by values in the "nodes" $\mathbf{N}_j$, $i = 1, \ldots, n$. Thus, a basis can be

$$\phi_j(\mathbf{N}_i) = \delta_{ij} = \begin{cases} 0 & i \neq j, \\ 1 & i = j \end{cases}$$



If $r = 1$, the nodes coincide with the triangle vertices (in the interior). [a.k.a. Lagrangian Finite Elements]

- $v_h \in V_h$ is then a linear combination of ϕ_i's:

$$v_h(x) = \sum_{i=1}^{n} v_i \phi_i(x) \qquad \forall x \in \Omega,$$

Galerkin method VI

Bases and the FEM

- v_i can be evaluations at the nodes $v_i = v_h(\mathbf{N}_i)$
- Back to Poisson

$$\int_\Omega \nabla u_h \cdot \nabla v_h \, dx = \int_\Omega f v_h \, dx \qquad \forall v_h \in V_h$$

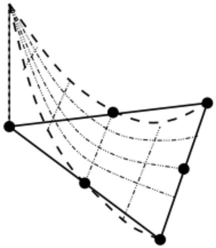
- Expanding also the discrete solution, the Galerkin method gives

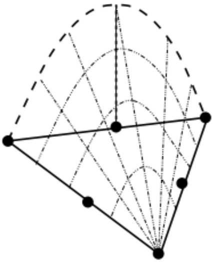
$$\sum_{j=1}^{n} u_j \int_\Omega \nabla \phi_j \cdot \nabla \phi_i \, dx = \int_\Omega f \, \phi_i \, dx, \qquad i = 1, \ldots, n$$

- Stiffness matrix $(n \times n)$ A with $a_{ij} = \int_\Omega \nabla \phi_j \cdot \nabla \phi_i \, dx$
- $A\boldsymbol{u} = \mathbf{f}$

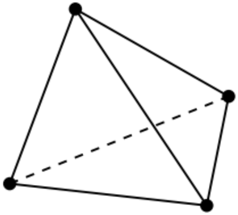
Other elements

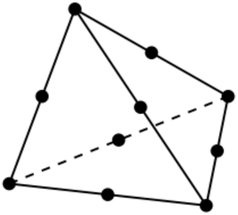
Higher-order





3-d





Other boundary conditions I: inhomoegeneous

Inhomogeneous Dirichlet b.c.

$$-\nabla^2 u = f \quad \text{in } \Omega \subset \mathbb{R}^d, \qquad u = g \text{ on } \partial\Omega.$$

- Take $u_h(x) = \sum_{j=1}^{n} c_j \phi_j(x) + \sum_{j=n+1}^{n+n_d} g(x_j)\phi_j(x)$
  - red term prescribed s.t. b.c. satisfied
  - e.g. $\phi_{n+\ell}(x)$ hat func at $x_{n+\ell} \in \partial\Omega$
- The rest remain same; note test space does not include $\phi_{n+\ell}$

Other boundary conditions II: Neumann

University of Oxford Mathematical Institute

Neumann b.c.

$$-\nabla^2 u = f \quad \text{in } \Omega \subset \mathbb{R}^d, \qquad \nabla u \cdot \mathbf{n} = g \text{ on } \partial\Omega.$$

Recall weak form

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_{\partial\Omega} (\nabla u \cdot \mathbf{n}) v \, ds + \int_\Omega f v \, dx = \int_{\partial\Omega} g v \, ds + \int_\Omega f v \, dx$$

- Take $u_h(x) = \sum_{j=1}^{n+n_e} c_j \phi_j(x)$ ($\phi_{n+\ell}(x)$ nonzero on $\partial\Omega$)
- test space $\xi_j = \phi_j$, $j = 1, \dots, n+n_e$
- Note $\int_{\partial\Omega} g v \, ds$ influences right-hand side in $A\mathbf{c} = \mathbf{f}$
- Robin ($u + \nabla u \cdot \mathbf{n} = g$ on $\partial\Omega$) or mixed ($u = g_1$ on $\partial\Omega$, $\nabla u \cdot \mathbf{n} = g_2$ on $\partial\Omega_2$) b.c. possible

Analysis of FEM I

Steps in estimating the error

1. Estimate the local interpolation error $v - \Pi_K^r v$, where

$$\Pi_K^r : C^0(K) \to \mathbb{P}_r(K), \qquad v \mapsto \Pi_K^r v$$

2. Extension of the estimate to the whole mesh

$$|v - \Pi_K^r v|_{m,\Omega} \leq C h^{r+1-m} |v|_{r+1,K}, \quad m = 0, 1$$

3. Error estimate in the "energy norm" (C indep. of h and u)

$$\|u - u_h\|_{1,\Omega} \leq C_1 C_2^{-1} h^r |u|_{r+1,\Omega}$$

Evidently, 2 ways of increase accuracy (reduce h or increase r). The latter effective only if u is smooth enough...

If $u \in H^{p+1}(\Omega)$ for some $p > 0$, then

$$\|u - u_h\|_{1,\Omega} \leq C h^s |u|_{s+1,\Omega}, \quad s = \min\{r, p\}$$

Then, if e.g. $u \in H^2(\Omega)$ (i.e. $p = 1$), then going for polynomials of degree ≥ 2 won't get you more accuracy

Summary:

r	$u \in H^1(I)$ ($p=0$)	$u \in H^2(I)$ ($p=1$)	$u \in H^3(I)$ ($p=2$)	$u \in H^4(I)$ ($p=3$)	$u \in H^5(I)$ ($p=4$)
1	converges	$\boxed{h^1}$	h^1	h^1	h^1
2	converges	h^1	$\boxed{h^2}$	h^2	h^2
3	converges	h^1	h^2	$\boxed{h^3}$	h^3
4	converges	h^1	h^2	h^3	$\boxed{h^4}$

Analysis of FEM III

Steps in estimating the error

Sometimes we're also interested in L^2-norm estimates. For Poisson one can prove that if $u \in H^{p+1}(\Omega)$ for some $p > 0$, then

$$\|u - u_h\|_{0,\Omega} \leq Ch^{s+1}|u|_{s+1,\Omega}, \quad s = \min\{r, p\}$$

Generalised Stokes equations

Strong form

We study the *generalised Stokes* problem with homogeneous Dirichlet boundary conditions

$$\begin{aligned} \mathbf{u} - \nu\Delta\mathbf{u} + \nabla p &= \mathbf{f} \quad \text{in } \Omega, \\ \nabla\cdot\mathbf{u} &= 0 \quad \text{in } \Omega, \\ \mathbf{u} &= \mathbf{0} \quad \text{on } \partial\Omega, \end{aligned}$$

- $\mathbf{u}$ vector field (in $\mathbb{R}^2$ or $\mathbb{R}^3$), p: pressure (scalar func.) the medium)
- describe the steady motion of an incompressible viscous fluid in a porous domain
- the model is valid for $Re \ll 1$

Generalised Stokes equations

Weak form

- Testing against $\mathbf{v}$,q, integrate over Ω, and apply IBP on the momentum equation: find $\mathbf{u} \in \boldsymbol{V}$ and $p \in Q_0$ (**mixed** FEM) st

$$\int_\Omega (\mathbf{u}\cdot\mathbf{v} + \nu\nabla\mathbf{u} : \nabla\mathbf{v}) - \int_\Omega p\nabla\cdot\mathbf{v} = \int_\Omega \mathbf{f}\cdot\mathbf{v} \quad \forall\mathbf{v}\in\boldsymbol{V},$$
$$\int_\Omega q\nabla\cdot\mathbf{u} = 0 \quad \forall q\in Q_0,$$

where $\boldsymbol{V} = [H_0^1(\Omega)]^d$ and $Q_0 = L_0^2(\Omega) = \{q\in L^2(\Omega) : q = 0 \text{ on } \partial\Omega\}$,
$\nabla\mathbf{u} : \nabla\mathbf{v} = \nabla u_x\cdot\nabla v_x + \nabla u_y\cdot\nabla v_y$ (in 2d)

- bilinear forms $a : \boldsymbol{V}\times\boldsymbol{V}\to\mathbb{R}$ and $b : \boldsymbol{V}\times Q\to\mathbb{R}$, and functional $\mathscr{F}(\mathbf{v}) = \int_\Omega \mathbf{f}\cdot\mathbf{v}$:

$$a(\mathbf{u},\mathbf{v}) = \int_\Omega (\mathbf{u}\cdot\mathbf{v} + \nu\nabla\mathbf{u}\cdot\nabla\mathbf{v}), \qquad b(\mathbf{u},q) = -\int_\Omega q\nabla\cdot\mathbf{u}.$$

- Find $(\mathbf{u},p)\in\boldsymbol{V}\times Q_0$ such that

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) &= \mathscr{F}(\mathbf{v}) && \forall\mathbf{v}\in\boldsymbol{V},\\ b(\mathbf{u},q) &= 0 && \forall q\in Q_0, \end{aligned}$$

Galerkin (conforming) finite element method I

- For Stokes eqn: find $(\mathbf{u}_h, p_h) \in \boldsymbol{V}_h \times Q_h$ such that

$$\begin{aligned} a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) &= \mathscr{F}(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \boldsymbol{V}_h, \\ b(\mathbf{u}_h, q_h) &= 0 \qquad\qquad \forall q_h \in Q_h, \end{aligned}$$

- $\{\boldsymbol{V}_h \subset \boldsymbol{V}\}$ and $\{Q_h \subset Q_0\}$ are families of finite dimensional subspaces

Galerkin (conforming) finite element method II

Mathematical Institute

Find $\mathbf{u} \in \boldsymbol{V}$ and $p \in Q_0$ (**mixed** FEM) st

$$\int_\Omega (\mathbf{u}\cdot\mathbf{v} + \nu \nabla \mathbf{u} : \nabla \mathbf{v}) - \int_\Omega p \nabla \cdot \mathbf{v} = \int_\Omega \mathbf{f}\cdot\mathbf{v} \quad \forall \mathbf{v} \in \boldsymbol{V},$$

$$\int_\Omega q \nabla \cdot \mathbf{u} = 0 \quad \forall q \in Q_0,$$

Associated linear system.

- $\{\varphi_j\}_{j=1}^N$ and $\{\phi_k\}_{k=1}^M$, basis functions for $\boldsymbol{V}_h$ and Q_h
- $\mathbf{u}_h = \sum_{j=1}^N u_j \varphi_j(x)$, $p_h = \sum_{k=1}^M p_k \phi_k(x)$, with $N = \dim(\boldsymbol{V}_h), M = \dim(Q_h)$
- Choosing the basis functions as tests:

$$\begin{aligned} \mathrm{A}\mathbf{U} + \mathrm{B}^T\mathbf{P} &= \mathbf{F}, \\ \mathrm{B}\mathbf{U} &= \mathbf{0}, \end{aligned} \quad \Leftrightarrow \quad \begin{pmatrix} \mathrm{A} & \mathrm{B}^T \\ \mathrm{B} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{U} \\ \mathbf{P} \end{pmatrix} = \begin{pmatrix} \mathbf{F} \\ \mathbf{0} \end{pmatrix},$$

Galerkin (conforming) finite element method III

UNIVERSITY OF OXFORD
Mathematical Institute

- $\mathrm{A} \in \mathbb{R}^{N\times N}$ and $\mathrm{B} \in \mathbb{R}^{M\times N}$ are associated to $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$

$$(\mathrm{A})_{ij} = a(\varphi_j, \varphi_i), \quad \mathrm{B}_{kj} = b(\varphi_j, \phi_k), \qquad i,j = 1,\ldots,N,\ k = 1,\ldots,M.$$

- Unknowns: $\mathbf{U} = (u_1,\ldots,u_N)^T, \quad \mathbf{P} = (p_1,\ldots,p_M)^T$
- Datum: $\mathbf{F} = (f_1,\ldots,f_N)^T$ with $f_i = \int_\Omega \mathbf{f}\cdot\varphi_i$
- The (generalised) Stokes matrix

$$\mathrm{S} = \begin{pmatrix} \mathrm{A} & \mathrm{B}^T \\ \mathrm{B} & 0 \end{pmatrix} \in \mathbb{R}^{(N+M)\times(N+M)}$$

  is block-symmetric (since A is symmetric) and indefinite (positive and negative eigenvalues)

- A stable solver is MINRES (symmetric variant of GMRES); preconditioning of course important

CG optimality from FEM arguments

$Ax_* = b$, x_k: CG solution after k steps, i.e., $Q^T(Ax_k - b) = 0$. Then

$$x_k = \operatorname{argmin}_{x\in \operatorname{span}(Q)}\|x - x_*\|_A$$

Since $\|y\|_A^2 = (y,y)_A = y^T Ay = \|A^{1/2}y\|^2$, statement equivalent to

$$\|A^{1/2}(x_k - x_*)\| = \min_x\{\|A^{1/2}(x - x_*)\| : x = \sum_{i=1}^k y_i q_i\}.$$

FEM-type proof: (recall Poisson) for any $y \in Q$,

$$\begin{aligned}\|A^{1/2}(x_k - x_*)\|^2 &= (x_k - x_*, x_k - x_*)_A = (x_k - x_*, x_k - y + y - x_*)\\ &= (x_k - x_*, y - x_*)_A + (x_k - x_*, x_k - y)_A\\ &= (x_k - x_*, y - x_*)_A\end{aligned}$$

due to Galerkin orthogonality:

$$(x_k - x_*, x_k - y)_A = (A(x_k - x_*), x_k - y) = (Ax_k - b, x_k - y) = (r, x_k - y) = 0.$$

By Cauchy-Schwarz,

$$\|A^{1/2}(x_k - x_*)\|^2 \le \|A^{1/2}(x_k - x_*)\|\|A^{1/2}(x_* - y)\|.$$

CG convergence

$A^T = A$ positive definite. Let $e_k := x_* - x_k$. $e_0 = x_*$, and

$$\begin{aligned}
\frac{\|e_k\|_A}{\|e_0\|_A} &= \min_{x \in \mathscr{K}_{k-1}(A,b)} \|x_k - x_*\|_A / \|x_*\|_A \\
&= \min_{p_{k-1} \in \mathscr{P}_{k-1}} \|p_{k-1}(A)b - A^{-1}b\|_A / \|e_0\|_A \\
&= \min_{p_{k-1} \in \mathscr{P}_{k-1}} \|(p_{k-1}(A)A - I)e_0\|_A / \|e_0\|_A \\
&= \min_{p \in \mathscr{P}_k, p(0)=1} \|p(A)e_0\|_A / \|e_0\|_A \\
&= \min_{p \in \mathscr{P}_k, \color{red}{p(0)=1}} \|Q \begin{bmatrix} p(\lambda_1) & & \\ & \ddots & \\ & & p(\lambda_n) \end{bmatrix} Q^T e_0\|_A / \|e_0\|_A
\end{aligned}$$

Now $\|Q \begin{bmatrix} p(\lambda_1) & & \\ & \ddots & \\ & & p(\lambda_n) \end{bmatrix} Q^T e_0\|_A^2 = \sum_i \lambda_i p(\lambda_i)^2 (Q^T e_0)_i \leq$
$\max_j p(\lambda_j)^2 \sum_i \lambda_i (Q^T e_0)_i = \max_j p(\lambda_j)^2 \|e_0\|_A^2$

CG convergence cont'd

University of Oxford Mathematical Institute

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq \|Q\|_A \|Q^T\|_A \min_{p \in \mathscr{P}_k, p(0)=1} \max |p(\lambda_i)|$$

Now

$$\min_{p \in \mathscr{P}_k, p(0)=1} \max |p(\lambda_i)| \leq \left(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1} \right)^k$$

- note $\kappa_2(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$
- obtained by Chebyshev polynomial on $[\lambda_{\min}(A), \lambda_{\max}(A)]$

MINRES convergence

(special case of GMRES) $A^T = A$ Recall that $x \in \mathscr{K}_k(A,b) \Rightarrow x = p_{k-1}(A)b$. Hence MINRES solution is

$$\begin{aligned}
\min_{x\in\mathscr{K}_k(A,b)} \|Ax-b\|_2 &= \min_{p_{k-1}\in\mathscr{P}_{k-1}} \|Ap_{k-1}(A)b-b\|_2 \\
&= \min_{\tilde{p}\in\mathscr{P}_k,\tilde{p}(0)=0} \|(\tilde{p}(A)-I)b\|_2 \\
&= \min_{p\in\mathscr{P}_k,p(0)=1} \|p(A)b\|_2
\end{aligned}$$

A is diagonalizable $A = Q\Lambda Q^T$, so

$$\begin{aligned}
\|p(A)\|_2 = \|Qp(\Lambda)Q^T\|_2 &\le \|Q\|_2\|Q^T\|_2\|p(\Lambda)\|_2 \\
&= \max_{z\in\lambda(A)} |p(z)|
\end{aligned}$$

Interpretation: (again) find polynomial s.t. $p(0) = 1$ and $|p(\lambda_i)|$ small

MINRES convergence cont'd

UNIVERSITY OF OXFORD
Mathematical Institute

$$\frac{\|Ax-b\|_2}{\|b\|_2} \leq \min_{p \in \mathscr{P}_k, p(0)=1} \max |p(\lambda_i)|$$

Now

$$\min_{p \in \mathscr{P}_k, p(0)=1} \max |p(\lambda_i)| \leq \left(2\frac{\kappa_2(A)-1}{\kappa_2(A)+1}\right)^{k/2}$$

- minimization needed on positive and negative sides, hence slower convergence when A indefinite (same bound as CG when $A \succ 0$)
- obtained by Chebyshev+change of variables [A. Greenbaum's book]

Navier-Stokes equation, very briefly

Steady-state Navier-Stokes equation

$$-\nu \nabla^2 \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \nabla p = \mathbf{f}$$

$$\nabla \cdot \mathbf{u} = 0$$

- Nonlinear in $\mathbf{u}$: iterative solution of linearized problems necessary (Picard, Newton)
- Multiple stable solutions can exist
- See e.g. Elman-Silvester-Wathen Ch.8

Backup slides

(from Ricardo)

Generalised Stokes equations

Choosing the spaces wisely, we can eliminate p

- Subspaces of $[H^1(\Omega)]^d$:

$$\boldsymbol{V}_{\mathrm{div}} = \{\mathbf{v} \in [H^1(\Omega)]^d \,:\, \nabla \cdot \mathbf{v} = 0 \text{ in } \Omega\}, \quad \boldsymbol{V}^0_{\mathrm{div}} = \{\mathbf{v} \in \boldsymbol{V}_{\mathrm{div}} \,:\, \mathbf{v} = \mathbf{0} \text{ on } \Gamma_D\}.$$

- Take $\mathbf{v} \in \boldsymbol{V}_{\mathrm{div}}$ in the momentum equation and the term involving the pressure p vanishes
- Equation only for the velocity:

$$\text{find } \mathbf{u} \in \boldsymbol{V}^0_{\mathrm{div}} : \quad a(\mathbf{u}, \mathbf{v}) = \int_\Omega \mathbf{f} \cdot \mathbf{v} \qquad \forall \mathbf{v} \in \boldsymbol{V}^0_{\mathrm{div}}.$$

- Well-posedness via Lax & Milgram
- Result: if we can solve the reduced problem in $\mathbf{u}$, then there exists a unique p st $(\boldsymbol{u}, p)$ is solution of the complete problem
- But! not practical since it requires to construct a FE space $\boldsymbol{V}_{\mathrm{div},h}$ of divergence-free functions (up to date, only 1 paper on that)
- Plus, how do I compute p?

Solvability theorem I

- Conditions for well-posedness:

  Abstract theory of saddle-point problems by Brezzi (1974)

Theorem: Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be Hilbert spaces. Consider $\mathscr{A}(\cdot,\cdot) : X \times X \to \mathbb{R}$, $\mathscr{B}(\cdot,\cdot) : X \times Y \to \mathbb{R}$, $\ell \in X'$, $\sigma \in Y'$, and the saddle-point problem: find $(u, \eta) \in X \times Y$ such that

$$\begin{aligned} \mathscr{A}(u,v) + \mathscr{B}(v,\eta) &= {}_{X'}\langle \ell, v\rangle_X \qquad \forall v \in X, & (2)\\ \mathscr{B}(u,\mu) &= {}_{Y'}\langle \sigma, \mu\rangle_Y \qquad \forall \mu \in Y. & (3) \end{aligned}$$

If the following hypotheses are satisfied:

1. $\mathscr{A}(\cdot,\cdot)$ is **continuous**: $|\mathscr{A}(u,v)| \leq \gamma \|u\|_X \|v\|_X \qquad \forall u, v \in X$
2. $\mathscr{A}$ is X^0-**elliptic**, with $X^0 = \{v \in X : \mathscr{B}(v,\mu) = 0 \; \forall \mu \in Y\}$,

$$|\mathscr{A}(v,v)| \geq \|v\|_X^2 \qquad \forall v \in X^0;$$

Solvability theorem II

University of Oxford
Mathematical Institute

3. $\mathscr{B}(\cdot,\cdot)$ is **continuous**: $|\mathscr{B}(u,\mu)| \leq \delta \|u\|_X \|\mu\|_Y \qquad \forall u \in X,\ \forall \mu \in Y$

4. **inf-sup condition**: $\exists \beta^* > 0$ st. $\displaystyle\inf_{\mu \in Y, \mu \neq 0} \sup_{v \in X, v \neq 0} \frac{\mathscr{B}(v,\mu)}{\|v\|_X \|\mu\|_Y} \geq \beta^*$

Then, (2)-(3) has a unique solution $(u,\eta) \in X \times Y$ and

$$
\begin{aligned}
\|u\|_X &\leq \left[\|\ell\|_{X'} + \frac{1+\gamma}{\beta^*} \|\sigma\|_{Y'} \right] \\
\|\eta\|_Y &\leq \frac{1}{\beta^*} \left[\left(1 + \frac{\gamma}{\bar{\alpha}}\right) \|\ell\|_{X'} + \frac{\gamma(\bar{\alpha}+\gamma)}{\bar{\alpha}\beta^*} \|\sigma\|_{Y'} \right].
\end{aligned}
$$

The Stokes equation falls in this framework with $X = \boldsymbol{V}$, $X^0 = \boldsymbol{V}_{\mathrm{div}}$, knowing that $H_0^1(\Omega)$ and $L^2(\Omega)$ satisfy the inf-sup condition

Galerkin (conforming) finite element method I

- For the Brinkman problem: find $(\mathbf{u}_h, p_h) \in \boldsymbol{V}_h \times Q_h$ such that

$$\begin{aligned} a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) &= \mathscr{F}(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \boldsymbol{V}_h, \\ b(\mathbf{u}_h, q_h) &= 0 \qquad \forall q_h \in Q_h, \end{aligned}$$

- $\{\boldsymbol{V}_h \subset \boldsymbol{V}\}$ and $\{Q_h \subset Q_0\}$ are families of finite dimensional subspaces

Galerkin (conforming) finite element method II

- Solvability also falls into the Brezzi theory with $X = \boldsymbol{V}_h$ and $X^0 = \boldsymbol{V}_h^0 = \{\mathbf{v}_h \in \boldsymbol{V}_h : b(\mathbf{v}_h, q_h) = 0\, \forall q_h \in Q_h\}$
- $\beta^* > 0$ appearing in the inf-sup condition may depend on h!

$$\exists \beta^* > 0 : \quad \inf_{q_h \in Q_h, q_h \neq 0} \sup_{\mathbf{v}_h \in \boldsymbol{V}_h, \mathbf{v}_h \neq \mathbf{0}} \frac{b(\mathbf{v}_h, q_h)}{\|\mathbf{v}_h\|_{H^1(\Omega)} \|q_h\|_{L^2(\Omega)}} \geq \beta^*$$

- A-priori estimates

$$\|\mathbf{u}_h\|_{\boldsymbol{V}} \leq \frac{1}{\bar{\alpha}} \|\mathbf{f}\|_{\boldsymbol{V}'}, \qquad \|p_h\|_Q \leq \frac{1}{\beta}\left(1 + \frac{\gamma}{\bar{\alpha}}\right) \|\mathbf{f}\|_{\boldsymbol{V}'},$$

- Céa's lemma

$$\|\mathbf{u} - \mathbf{u}_h\|_{\boldsymbol{V}} \leq \left(1 + \frac{\gamma}{\beta^*}\right)\left(1 + \frac{\gamma}{\bar{\alpha}}\right) \inf_{\mathbf{v}_h \in \boldsymbol{V}_h} \|\mathbf{u} - \mathbf{v}_h\|_{\boldsymbol{V}} + \frac{\delta}{\bar{\alpha}} \inf_{q_h \in Q_h} \|p - q_h\|_Q,$$

$$\|p - p_h\|_Q \leq \frac{\gamma}{\beta^*}\left(1 + \frac{\gamma}{\bar{\alpha}}\right) \inf_{\mathbf{v}_h \in \boldsymbol{V}_h} \|\mathbf{u} - \mathbf{v}_h\|_{\boldsymbol{V}} + \left(1 + \frac{\delta}{\beta^*} + \frac{\delta\gamma}{\bar{\alpha}\beta^*}\right) \inf_{q_h \in Q_h} \|p - q_h\|_Q.$$

Associated linear system.

- $\{\varphi_j\}_{j=1}^N$ and $\{\phi_k\}_{k=1}^M$, basis functions for $\boldsymbol{V}_h$ and Q_h
- $\mathbf{u}_h = \sum_{j=1}^N u_j \varphi_j(x)$, $p_h = \sum_{k=1}^M p_k \phi_k(x)$, with $N = \dim(\boldsymbol{V}_h), M = \dim(Q_h)$
- Choosing the basis functions as tests:

$$\begin{aligned} \mathrm{A}\mathbf{U} + \mathrm{B}^T\mathbf{P} &= \mathbf{F}, \\ \mathrm{B}\mathbf{U} &= \mathbf{0}, \end{aligned} \quad \Leftrightarrow \quad \begin{pmatrix} \mathrm{A} & \mathrm{B}^T \\ \mathrm{B} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{U} \\ \mathbf{P} \end{pmatrix} = \begin{pmatrix} \mathbf{F} \\ \mathbf{0} \end{pmatrix},$$

Galerkin (conforming) finite element method IV

- $\mathrm{A} \in \mathbb{R}^{N\times N}$ and $\mathrm{B} \in \mathbb{R}^{M\times N}$ are associated to $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$

$$(\mathrm{A})_{ij} = a(\varphi_j, \varphi_i), \quad \mathrm{B}_{kj} = b(\varphi_j, \phi_k), \qquad i,j = 1,\ldots,N,\ k = 1,\ldots,M.$$

- Unknowns: $\mathbf{U} = (u_1,\ldots,u_N)^T, \quad \mathbf{P} = (p_1,\ldots,p_M)^T$
- Datum: $\mathbf{F} = (f_1,\ldots,f_N)^T$ with $f_i = \int_\Omega \mathbf{f}\cdot\varphi_i$
- The (generalised) Stokes matrix

$$\mathrm{S} = \begin{pmatrix} \mathrm{A} & \mathrm{B}^T \\ \mathrm{B} & 0 \end{pmatrix} \in \mathbb{R}^{(N+M)\times(N+M)}$$

is block-symmetric (since A is symmetric) and non-definite (real eigenvalues of variable sign)

More on the discrete inf-sup condition I

- The algebraic problem has a unique solution iff $det(\mathbb{S}) \neq 0$ (true if the discrete inf-sup condition holds)
- If the inf-sup condition is not satisfied

$$\exists q_h^* \in Q_h : \quad b(\mathbf{v}_h, q_h^*) = 0 \qquad \forall \mathbf{v}_h \in \boldsymbol{V}_h .$$

- Thus, if $(\boldsymbol{u}_h, p_h)$ is a solution, then also $(\mathbf{u}_h, p_h + q_h^*)$, because $a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h + q_h^*) = a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) + \underbrace{b(\mathbf{v}_h, q_h^*)}_{=0} =$

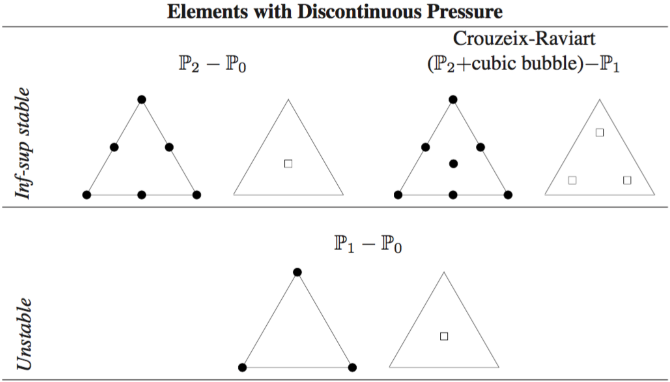
  $\mathscr{F}(\mathbf{v}_h), \ \forall \mathbf{v}_h \in \boldsymbol{V}_h$.
- Non-uniqueness!!

More on the discrete inf-sup condition II

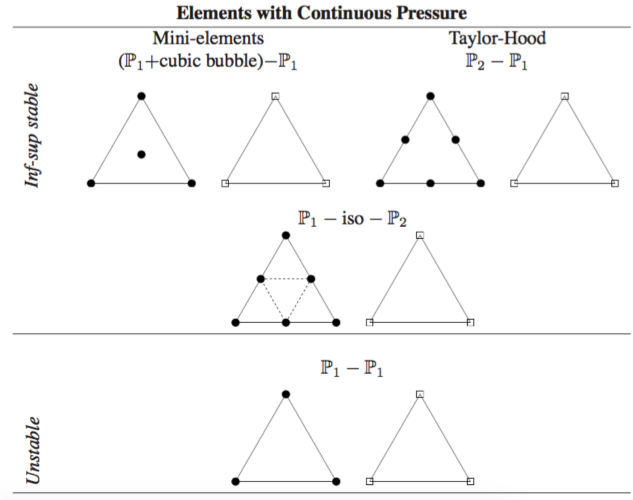
- p_h^* breaking the inf-sup condition are called spurious pressure modes
- Who's fault is this?!! Q_h and $\boldsymbol{V}_h$ ...
- Pairs $(\boldsymbol{V}_h, Q_h)$ violating the inf-sup condition are called inf-sup unstable
- The weak form does not require the pressure to be continuous
- Possible choices (degrees of freedom of the velocity "$\bullet$" and those of the pressure are "$\square$")
- See a list in Girault-Raviart or Brezzi-Fortin books

More on the discrete inf-sup condition III

University of Oxford
Mathematical Institute



More on the discrete inf-sup condition IV



Stabilised formulations I

- Hope in the horizon: you can still use unstable pairs (why would you want to do that?)
- Some remedies available (cf Exercises of week 4)
- General stabilisation technique: find $\boldsymbol{u}_h \in \boldsymbol{V}_h$, $q_h \in Q_h$ such that

$$\begin{aligned} a(\boldsymbol{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) &= \mathscr{F}(\mathbf{v}_h) - \Psi_h^{(\rho)}(\mathbf{v}_h) \quad \forall \mathbf{v}_h \in \boldsymbol{V}_h \\ b(\boldsymbol{u}_h, q_h) &= \Phi_h(q_h) \quad \forall q_h \in Q_h, \end{aligned}$$

where

$$\begin{aligned} \Psi_h^{(\rho)}(\mathbf{v}_h) &= \bar{\delta} \sum_{K \in \mathscr{T}_h} h_K^2 \int_K (\alpha \boldsymbol{u}_h - \nu \triangle \boldsymbol{u}_h + \nabla p_h - \mathbf{f}) \cdot (\rho \alpha \mathbf{v}_h - \rho \nu \triangle \mathbf{v}_h) \\ \Phi_h(q_h) &= \bar{\delta} \sum_{K \in \mathscr{T}_h} h_K^2 \int_K (\alpha \boldsymbol{u}_h - \nu \triangle \boldsymbol{u}_h + \nabla p_h - \mathbf{f}) \cdot \nabla q_h . \end{aligned}$$

with $\bar{\delta} > 0, \rho$ stabilisation parameters to be set

Stabilised formulations II

- $\rho = 0 \Rightarrow \Psi_h^{(0)} = 0 \leftrightarrow$ Streamline Upwind/Petrov-Galerkin (SUPG) method
- $\rho = -1 \leftrightarrow$ Galerkin/Least-Squares (GLS or GaLS) method
- These methods are strongly consistent (other versions may not)

Stokes flow ($\alpha = 0$).

- Notice that if using $\mathbb{P}_1 - \mathbb{P}_1$ elements, then $\Delta \mathbf{v}_h = \Delta \boldsymbol{u}_h = \mathbf{0}$ for all $K \in \mathscr{T}_h$
- The stabilised method is well-posed for adequate stabilisation parameters (see e.g. Quarteroni-Valli, section 9.4)
- Stability and convergence also follow

Stabilised formulations III

- Matrix form

$$\begin{pmatrix} \mathrm{A} & \mathrm{B}^T \\ \mathrm{B} & -\mathrm{C} \end{pmatrix} \begin{pmatrix} \mathbf{U} \\ \mathbf{P} \end{pmatrix} = \begin{pmatrix} \mathbf{F} \\ \mathbf{G} \end{pmatrix}$$

with $\mathrm{C}_{km} = \bar{\delta} \sum_{K \in \mathscr{T}_h} h_K^2 \int_K \nabla \phi_m \cdot \nabla \phi_k, \qquad k, m = 1, \ldots, M$

$$g_k = -\bar{\delta} \sum_{K \in \mathscr{T}_h} h_K^2 \int_K \mathbf{f} \cdot \nabla \phi_k, \qquad k = 1, \ldots, M.$$

- Similar method (also with a "name"): Brezzi-Pitkaranta (uses $\mathbb{P}_1 - \mathbb{P}_1$)

$$\begin{aligned} a_0(\boldsymbol{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) &= \mathscr{F}(\mathbf{v}_h) && \forall \mathbf{v}_h \in \boldsymbol{V}_h \\ b(\boldsymbol{u}_h, q_h) &= \textstyle\sum_{K \in \mathscr{T}_h} \delta_K (\nabla p_h, \nabla q_h)_{0,K} && \forall q_h \in Q_h, \end{aligned}$$

with $\delta_K = \dfrac{|K|^2}{5(c_1^2 + c_2^2 + c_3^2)}$, $|K|$: area of K, c_i : length of edges