The use of stopping criteria for iterative Krylov methods in designing adaptive methods for PDEs

Mario Arioli

Visiting Professor, Wuppertal University

January 6, 2016

Collaborations

E. Georgoulis, J. Liesen, D. Loghin, A.Miedlar, E. Noulard, D. Orban, A. Russo, Z. Strakos, and A. Wathen

Problem

$H^1_0(\omega)$ the standard Sobolev space of functions with zero trace on $\partial\omega$.
Let Ω be a bounded open polyhedral domain in $\mathbb{R}^d$, $d = 2, 3$ and let $\partial\Omega$ denote its boundary. We consider the second order equation

$$(\clubsuit) \qquad -\nabla \cdot (a\nabla u) = f \quad \text{in } \Omega,$$

where $a \in [L^\infty(\Omega)]^{d\times d}$ is a positive definite tensor and $f \in L^2(\Omega)$.
For simplicity of the presentation, we impose homogeneous Dirichlet boundary condition $u = 0$ on $\partial\Omega$, although this appears not to be an essential restriction. We shall denote by $||\cdot||_a := \|\sqrt{a}\nabla(\cdot)\|$ the, so-called, energy norm.

FEM

Let $\mathcal{T}$ be a conforming subdivision of Ω into disjoint simplicial elements $\kappa \in \mathcal{T}$. We assume that the subdivision $\mathcal{T}$ is shape-regular and that it is constructed via affine mappings F_κ, where $F_\kappa : \hat{\kappa} \to \kappa$, with non-singular Jacobian, where $\hat{\kappa}$ is the reference simplex.
For a nonnegative integer r, we denote by $\mathcal{P}_r(\hat{\kappa})$, the set of all polynomials of total degree at most r, defined on $\hat{\kappa}$. We consider the finite element space

$$\mathbb{V} := \{V \in H_0^1(\Omega) : \; V|_\kappa \circ F_\kappa \in \mathcal{P}_r(\hat{\kappa}), \, \kappa \in \mathcal{T}\}.$$

FEM

By Γ we denote the union of all $(d-1)$-dimensional element faces associated with the subdivision $\mathcal{T}$ (including the boundary). Further we decompose Γ into two disjoint subsets $\Gamma = \partial\Omega \cup \Gamma_{\mathrm{int}}$, where $\Gamma_{\mathrm{int}} := \Gamma \backslash \partial\Omega$. We define $h_\kappa := (\mu_d(\kappa))^{1/d}$, $\kappa \in \mathcal{T}$, where μ_d is the d-dimensional Lebesgue measure. Also, for two (generic) elements κ^+, κ^- sharing a face $e := \partial\kappa^+ \cap \partial\kappa^- \subset \Gamma_{\mathrm{int}}$ we define $h_e := \mu_{d-1}(e)$. We collect these quantities into the element-wise constant function $\mathbf{h} : \Omega \to \mathbb{R}$, with $\mathbf{h}|_\kappa = h_\kappa$, $\kappa \in \mathcal{T}$ and $\mathbf{h}|_e = h_e$, $e \in \Gamma$. The families of meshes constructed by the algorithms presented in this work will be conforming and shape-regular.

FEM

The finite element method reads:

$$(\bigstar) \qquad \text{find } U \in \mathbb{V} \text{ such that} \quad a(U, V) = l(V) \quad \forall V \in \mathbb{V},$$

where the bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ and the linear form $l : H_0^1(\Omega) \to \mathbb{R}$ are given by

$$a(w, v) := \int_\Omega a \nabla w \cdot \nabla v \, \mathrm{d}x \qquad \text{and} \qquad l(v) := \int_\Omega f v \, \mathrm{d}x,$$

respectively, for $w, v \in H_0^1(\Omega)$.

FEM

Let now $\{\phi_i\}_{1\leq i\leq N}$ denote a set of basis functions for $\mathbb{V}$ so that

$$U = \sum_{i=1}^{N} \mathbf{u}_i \phi_i,$$

and let $\mathbf{A}_{ij} = a(\phi_j, \phi_i), \quad \mathbf{b}_k = l(\phi_k), \quad i, j, k = 1, \cdots, N$. With this notation, the linear system corresponding to is

$$\mathbf{A}\mathbf{u} = \mathbf{b},$$

where $\mathbf{A} \in \mathrm{R}^{N\times N}$ is the stiffness matrix corresponding to a set of basis functions $\{\phi_i\}_{1\leq i\leq N}$.

AFEM

For every face $e \in \Gamma_{\mathrm{int}}$, we define the *jump* across e of a scalar function w, defined in an open neighbourhood of e, by

$$[w](x) = \lim_{t \to 0} \Big(w(x - t\mathbf{n}_e) - w(x + t\mathbf{n}_e) \Big),$$

for $x \in e$, where $\mathbf{n}_e$ denotes a normal vector to e. (Note that the jump is only uniquely defined up to a sign, which is unimportant for the discussion below.) For any subset $\mathcal{M} \subset \mathcal{T}$ (i.e., $\mathcal{M}$ is a collection of elements of $\mathcal{T}$), we define the local estimator by

$$\eta_{\mathcal{T}}(U, \mathcal{M}) := \Big(\sum_{\kappa \in \mathcal{M}} \Big(h_\kappa^2 \| f + \nabla \cdot (a \nabla U) \|_\kappa^2 + \sum_{e \in \Gamma_{\mathrm{int}} \cap \partial \kappa} h_e \| [a \nabla U \cdot \mathbf{n}_e] \|_e^2 \Big) \Big)^{1/2}$$

AFEM

Algorithm 1. AFEM algorithm

```
Set parameter $0 < \theta \leq 1$. Set $m = 0$.
While convergence criterion not satisfied
    1. Solve exactly (★) to obtain $U_m^e$ (the exact solution).
    2. Compute local estimators $\eta_{\mathcal{T}_m}(U_m^e, \kappa)$, $\kappa \in \mathcal{T}_m$.
    3. Mark elements $\mathcal{M}_m$ for refinement in $\mathcal{T}_m$ using (Dörfler marking)
           $\eta^2_{\mathcal{T}_m}(U_m^e, \mathcal{M}_m) \geq \theta\ \eta^2_{\mathcal{T}_m}(U_m^e, \mathcal{T}_m)$.
    4. Refine $\mathcal{M}_m$ to obtain new mesh $\mathcal{T}_{m+1}$. Set $m \leftarrow m + 1$.
End
```

AFEM

$$\mathrm{SOLVE} \to \mathrm{ESTIMATE} \to \mathrm{MARK} \to \mathrm{REFINE}$$

Theorem

There exist constants $\xi > 0$ *and* $0 < \alpha < 1$ *such that*

$$\|u - U^e_{m+1}\|_a^2 + \xi \eta^2_{\mathcal{T}_{m+1}}(U^e_{m+1}, \mathcal{T}_{m+1}) \leq \alpha \left(\|u - U^e_m\|_a^2 + \xi \eta^2_{\mathcal{T}_m}(U^e_m, \mathcal{T}_m) \right).$$

(Cascon, Kreuzer, Nochetto, and Siebert SINUM 2008)

AFEM$\longmapsto$ iAFEM

SOLVE → ESTIMATE → MARK → REFINE

AFEM$\longmapsto$ iAFEM

APPROXIMATE → ESTIMATE → MARK → REFINE

AFEM$\longmapsto$ iAFEM

Algorithm 2. Inexact AFEM

```
Set parameters $0 < \theta \leq 1$, $\mu$ and $\nu$. Initialise $\tilde{U}_0$. Set $m = 1$.
While convergence criterion not satisfied
    1. Solve inexactly ($\bigstar$) to obtain $\tilde{U}_m$ so that
           $\|\tilde{U}_{m-1} - U_{m-1}\|^2 + \mu\|\tilde{U}_m - U_m\|^2 \leq \nu\eta^2_{m-1}(\tilde{U}_{m-1})$,
       for some values $\mu$ and $\nu$ is satisfied.
    2. Compute local estimators $\eta_{\tilde{\mathcal{T}}_m}(\tilde{U}_m, \kappa)$, $\kappa \in \tilde{\mathcal{T}}_m$.
    3. Mark elements $\tilde{\mathcal{M}}_m$ for refinement in $\tilde{\mathcal{T}}_m$ using
           $\eta^2_{\mathcal{T}_m}(\tilde{U}_m, \mathcal{M}_m) \geq \theta\ \eta^2_{\mathcal{T}_m}(\tilde{U}_m, \mathcal{T}_m)$.
    4. Refine $\tilde{\mathcal{M}}_m$ to obtain new mesh $\tilde{\mathcal{T}}_{m+1}$. Set $m \leftarrow m + 1$.
End
```

AFEM$\longmapsto$ iAFEM

Theorem

Let u, $\tilde{U}_m$ and $\tilde{U}_{m+1}$, $m \geq 1$ (approximations of U_m and U_{m+1} solutions on $\tilde{\mathcal{T}}_m$ and $\tilde{\mathcal{T}}_{m+1}$) , be such that

$$||\tilde{U}_m - U_m||_a^2 + \mu||\tilde{U}_{m+1} - U_{m+1}||_a^2 \leq \nu\eta_m^2(\tilde{U}_m),$$

with

$$\mu := \frac{1+\xi C_1(1+\gamma^{-1})}{\epsilon\xi\left(1+2C_2\right)}, \quad \nu := \frac{\beta}{\epsilon\left(1+2C_2C_1\right)},$$

where $0 < \epsilon < 1$, $\xi := \left(2C_1(1+\gamma)(1+\delta^{-1})\right)^{-1}$, and β, γ, δ and ϵ are chosen small enough, so that $(1-\tau\theta)(1+\delta) + 2\epsilon C_2 + \beta < 1$. Then, there exist a constant $0 < \alpha < 1$, depending only on the shape regularity of $\tilde{\mathcal{T}}_1$ and on the marking parameter θ, such that

$$||u - \tilde{U}_{m+1}||_a^2 + \xi\eta_{m+1}^2(\tilde{U}_{m+1}) \leq \alpha\left(||u - \tilde{U}_m||_a^2 + \xi\eta_m^2(\tilde{U}_m)\right).$$

(A., Georgoulis, and Loghin SISC, 2013)

AFEM$\longmapsto$ iAFEM

$$||\tilde{U}_m - U_m||_a^2 + \mu||\tilde{U}_{m+1} - U_{m+1}||_a^2 \leq \nu\eta_m^2(\tilde{U}_m),$$

AFEM$\longmapsto$ iAFEM

$$||\tilde{U}_m - U_m||_a^2$$

AFEM$\longmapsto$ iAFEM

$$\mathbf{A}_m \mathbf{u}_m = \mathbf{b}_m.$$

The matrices $\mathbf{A}_m \in \mathrm{R}^{N_m \times N_m}$ with $\{N_m\}_m$ an increasing sequence.

$$\| U_m - U_m^k \|_a = \| \mathbf{u}_m - \mathbf{u}_m^k \|_{\mathbf{A}_m}$$

where $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} := \mathbf{x}^T \mathbf{A} \mathbf{y}$, $\mathbf{x}, \mathbf{y} \in \mathrm{R}^N$, $\mathbf{A} \in \mathbb{R}^{N \times N}$, denotes the standard inner product weighted by $\mathbf{A}$ in $\mathbb{R}^N$, with the corresponding norm $\| \mathbf{x} \|_{\mathbf{A}} := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}}}$.

Krylov + CG

- We need a method that includes an energy-norm estimator (possibly an upper bound) of the errors!
- It would be desirable to have a monotonic sequence!

Krylov + CG

Let $\mathbf{u}_m^k \in \mathrm{R}^{N_m}$ be the k-th CG iterate at step m of the adaptive algorithm and by U_m^k the corresponding function in $\tilde{\mathbb{V}}_m$. We denote the residual by $\mathbf{r}_m^k := \mathbf{b}_m - \mathbf{A}_m \mathbf{u}_m^k$ and note that the energy norm of the error can be expressed as a dual norm of the residual:

$$\| U_m - U_m^k \|_a = \| \mathbf{u}_m - \mathbf{u}_m^k \|_{\mathbf{A}_m} = \| \mathbf{r}_m^k \|_{\mathbf{A}_m^{-1}},$$

Krylov + CG

It is well-known that the Conjugate Gradient method minimises the energy norm of the error, namely

$$\mathbf{u}_m^k = \arg \min_{\mathbf{u} \in \mathcal{K}_k(\mathbf{r}_m^0, \mathbf{A}_m)} \|\mathbf{u}_m - \mathbf{u}\|_{\mathbf{A}_m},$$

where $\mathcal{K}_k(\mathbf{r}_m^0, \mathbf{A}_m) := \left\{ \mathbf{r}_m^0, \mathbf{A}_m \mathbf{r}_m^0, \cdots, \mathbf{A}_m^{k-1} \mathbf{r}_m^0 \right\}$ is the Krylov subspace of dimension k. Thus, the energy norm of the error decreases monotonically and the criterion needed will be satisfied for all U_m^k with $k > k^*$ for some k^*.

In addition, there are various established numerical techniques that provide bounds or estimates for the energy norm of the error at each step.

Krylov + CG

We note that these properties do not hold in general, and that for non-symmetric problems, the best choice of iterative method remains unclear.

Error bounds for CG method

Algorithm 3. Conjugate Gradient Algorithm

```
Set $\mathbf{u}^0 := 0; \mathbf{p}^0 := \mathbf{r}^0 := \mathbf{b}; \sigma_0 := \|\mathbf{r}^0\|^2;$
For $j = 0, 1, \ldots$ until convergence do
    $\mathbf{v}^j = \mathbf{A}\mathbf{p}^j;\ \gamma_j = \sigma_j / (\mathbf{r}^j \cdot \mathbf{v}^j);$
    $\mathbf{u}^{j+1} = \mathbf{u}^j + \gamma_j \mathbf{p}^j;\ \mathbf{r}^{j+1} = \mathbf{r}^j - \gamma_j \mathbf{u}^j;\ \sigma_{j+1} = \|\mathbf{r}^{j+1}\|^2;$
    $\chi_{j+1} = \sigma_{j+1} / \sigma_j;\ \mathbf{p}^{j+1} = \mathbf{r}^{j+1} + \chi_{j+1} \mathbf{p}^j;$
End
```

Error bounds for CG method

The above algorithm constructs implicitly a Lanczos tridiagonalisation

$$\mathbf{V}_k^T \mathbf{A} \mathbf{V}_k = \mathbf{T}_k,$$

where $\mathbf{V}_k^T \mathbf{V} = \mathbf{I}_k$ and $\mathbf{T}_k \in \mathrm{R}^{k \times k}$ s.t.

$$\mathbf{T}_k = \begin{pmatrix} \alpha_1 & \beta_1 & & & 0 \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & & \alpha_{k-1} & \beta_{k-1} \\ 0 & & & \beta_{k-1} & \alpha_k \end{pmatrix}.$$

where for $j = 1, \ldots, k$,

$$\alpha_j = \frac{1}{\gamma_{j-1}} + \frac{\chi_{j-1}}{\gamma_{j-2}}, \quad \beta_j = \frac{\sqrt{\chi_j}}{\left|\gamma_{j-1}\right|},$$

with $\gamma_{-1} = 1, \chi_0 = 0$.

Error bounds for CG method: Hestenes and Stiefel

Hestenes-Stiefel rule (1952)

$$e_{\mathbf{A}}^{(k)} = \|\mathbf{u} - \mathbf{u}^k\|_{\mathbf{A}}^2 = \sum_{k+1}^{N} \gamma_j \|\mathbf{r}^j\|^2.$$

Error bounds for CG method: Hestenes and Stiefel

Under the assumption that $e_{\mathbf{A}}^{(k+d)} << e_{\mathbf{A}}^{(k)}$, where the integer d denotes a suitable delay, the Hestenes and Stiefel estimate is given by the formula (see A. 2003, Strakoš and Tichý, 2002)

$$\|\mathbf{u} - \mathbf{u}^k\|_A^2 \approx \sum_{j=k+1}^{k+d} \gamma_j \|\mathbf{r}^j\|^2.$$

$d = 10$ is indicated as a successful compromise, and numerical experiments support this conclusion (Golub and Meurant 97, A. 2004, and Golub-Meurant *Matrices, Moments and Quadrature with Applications, 2010*. However, numerical experiments indicate that the cheaper choice $d = 5$ can be reliable if the solution u is reasonably regular; in general, one can expect d to be required to be larger for ill-conditioned problems. Strakoš and Tichý, 2002 proved that it is numerically stable

Preconditioning

Let **B** a non singular matrix: the symmetric preconditioned system is

$$\mathbf{B}^{-T}\mathbf{A}\mathbf{B}^{-1}\mathbf{y} = \mathbf{B}^{-T}\mathbf{b} \qquad \left(\mathbf{y} = \mathbf{B}\mathbf{u}\right)$$

The dual norm of the preconditioned residual is equal to the dual norm of the original residual; i.e. the energy norm is “preconditioning invariant” for H-S.

The Golub and Meurant bounds

The **A**-norm of the error at each CG step can be written in the following way, using the orthogonality $\mathbf{r}_k^T \mathbf{u}^k = 0$,

$$\|\mathbf{u} - \mathbf{u}^k\|_{\mathbf{A}}^2 = \|\mathbf{r}_k\|_{\mathbf{A}^{-1}}^2 = \mathbf{b}^T \mathbf{A}^{-1} \mathbf{b} - \mathbf{b}^T \mathbf{u}^k.$$

Thus, the main difficulty in evaluating the above quantity is in the evaluation of the first term on the right-hand side. This term can be written as

$$F(\mathbf{A}) = \mathbf{b}^T \mathbf{A}^{-1} \mathbf{b} = \int_{\lambda_{\min}(\mathbf{A})}^{\lambda_{\max}(\mathbf{A})} \lambda^{-1} \mathrm{d}\omega(\lambda),$$

where the measure ω is a non-decreasing step function with jump discontinuities depending on the Fourier coefficients of **b** at the eigenvalues of **A**. Golub and Meurant used this formulation to provide upper and lower bounds on the CG error, by employing Gauss, Gauss-Radau and Gauss-Lobatto quadrature rules,

The Golub and Meurant bounds

The Gauss quadrature approach can be shown to be equivalent to the Hestenes and Stiefel estimate above.

The Golub and Meurant bounds

The only guaranteed upper bound for the $\mathbf{A}$-norm of the CG error uses a Gauss-Radau quadrature associated with the measure ω and with one node prescribed at $\lambda < \lambda_{\text{min}}(\mathbf{A})$.

The Golub and Meurant bounds

The only guaranteed upper bound for the **A**-norm of the CG error uses a Gauss-Radau quadrature associated with the measure ω and with one node prescribed at $\lambda < \lambda_{\min}(\mathbf{A})$. Let

$$\hat{T}_{k+1} = \begin{pmatrix} \alpha_1 & \beta_1 & & & 0 \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & & \alpha_k & \beta_k \\ 0 & & & \beta_k & \hat{\alpha}_{k+1} \end{pmatrix}.$$

where

$$\hat{\alpha}_{k+1} = \lambda + \beta_k^2 \mathbf{e}_k^T (T_k - \lambda I_k)^{-1} \mathbf{e}_k$$

with $\mathbf{e}_k$ the k-th column of the $k \times k$ identity matrix.

The Golub and Meurant bounds

The only guaranteed upper bound for the **A**-norm of the CG error uses a Gauss-Radau quadrature associated with the measure ω and with one node prescribed at $\lambda < \lambda_{\min}(\mathbf{A})$. Assuming $0 < \lambda < \lambda_{\min}(\mathbf{A})$, the Cholesky decomposition $\hat{\mathbf{T}}_{k+1} = \hat{\mathbf{R}}_{k+1}^T \hat{\mathbf{R}}_{k+1}$ can be shown to exist. Let now $\hat{\mathbf{y}}^{k+1}$ be the solution of

$$\hat{\mathbf{R}}_{k+1}^T \hat{\mathbf{y}}^{k+1} = \|\mathbf{b}\| \hat{\mathbf{e}}_1,$$

where $\hat{\mathbf{e}}_1$ denotes the first column of the identity matrix of size $k+1$. Then an upper bound on the CG error is given by

$$\|\mathbf{u} - \mathbf{u}_k\|_{\mathbf{A}} \leq \left| \hat{\mathbf{y}}_{k+1}^{k+1} \right| .$$

The Golub and Meurant bounds

It is clear that in order to compute this bound, the lower bound λ is required. In fact, experiments show that a close lower bound on the smallest eigenvalue of $\mathbf{A}$ yields tight upper bounds for the CG error (Golub-Meurant, 2010, A. -Georgoulis-Loghin, 2013).

λ and $\lambda_{\min}(\mathbf{A})$ depend on the preconditioning!

Adaptive stopping criteria for CG

Criterion

$$||\tilde{U}_m - U_m||_a^2 + \mu||\tilde{U}_{m+1} - U_{m+1}||_a^2 \leq \nu\eta_m^2(\tilde{U}_m),$$

cannot be employed in a practical context. Instead, the following generic criteria will be considered:

$$E(\tilde{U}_m)^2 + \mu E(\tilde{U}_{m+1})^2 \leq \nu\eta_m^2(\tilde{U}_m),$$

where $E(\tilde{U}_m)$ denotes an estimate or bound for the error $\|U_m - \tilde{U}_m\|_a$. Note that if $E(\tilde{U}_m)$ is an upper bound, then the result of the convergence Theorem hold and the inexact AFEM algorithm is guaranteed to converge. In general, estimates will not provide this guarantee, though a tight estimate or lower bound could also ensure the contraction result of the convergence Theorem, possibly at a different rate. For such cases, further analysis is required.

Test Problem 3D

Problem (♣) with $a = 1$ in $\Omega = (-1,\ 1)^3$ and the forcing function chosen so that the exact solution is $u = e^{-10r^2}$. We used the same Dörfler parameter $\theta = 0.75$ and started the adaptive algorithm from a range of initial regular meshes of tetrahedra and ran the procedure for $m = 10$ iterations. The refinement is concentrated near the origin, where the solution exhibits a sharp exponential decay.

Numerical experiments

We can estimate λ by

- Eigenvalue bounds based on Poincaré inequalities.
- Estimates using the Lanczos algorithm.

and then we can compute

$$\|\mathbf{u} - \mathbf{u}_k\|_{\mathbf{A}} \leq \left|\hat{\mathbf{y}}_{k+1}^{k+1}\right|.$$

Numerical experiments

We can estimate λ by

- Eigenvalue bounds based on Poincaré inequalities.
- Estimates using the Lanczos algorithm.

and then we can compute

$$\|\mathbf{u} - \mathbf{u}_k\|_{\mathbf{A}} \leq \left| \hat{\mathbf{y}}_{k+1}^{k+1} \right| .$$

1. DNR: the ideal bound using the exact dual norm of the residual;
2. GM1: the Golub-Meurant upper bound with adaptive bounds based on Poincaré for $\lambda_{\min}(\mathbf{A}_m)$;
3. GM2: the Golub-Meurant upper bound with global Poincaré bound for $\lambda_{\min}(\mathbf{A}_m)$;
4. GM3: the Golub-Meurant criterion with the Lanczos based estimator for $\lambda_{\min}(\mathbf{A}_m)$ with $c = 1/2$;
5. HS: the Hestenes-Stiefel estimator with a delay of $d = 5$ steps.
6. ER($|\log tol|$): the standard Euclidean residual with various stopping tolerances tol.

Selected experiments

	$N_0 = 142$			$N_0 = 779$			$N_0 = 5,191$		
method	N_m	$\|u - \tilde{U}_m\|_a$	mv	N_m	$\|u - \tilde{U}_m\|_a$	mv	N_m	$\|u - \tilde{U}_m\|_a$	mv
exact	19,579	8.9670e-2	–	131,250	4.7497e-2	–	950,961	†	–
DNR	19,507	8.9617e-2	120	131,452	4.7744e-2	302	†	†	†
GM1	19,573	8.9665e-2	179	131,243	4.7498e-2	447	951,057	2.4695e-2	1,065
GM2	19,582	8.9672e-2	250	131,232	4.7497e-2	570	950,988	2.4696e-2	1,292
GM3	19,510	8.9682e-2	151	131,251	4.7484e-2	353	951,077	2.4695e-2	1,018
HS	19,648	9.0596e-2	120	131,606	4.9477e-2	291	958,982	2.6542e-2	676
ER(6)	19,587	8.9677e-2	238	131,246	4.7497e-2	465	951,239	2.4692e-2	958
ER(8)	19,579	8.9670e-2	331	131,250	4.7497e-2	649	950,954	2.4697e-2	1,505
ER(10)	19,579	8.9670e-2	412	131,250	4.7497e-2	840	950,932	2.4697e-2	2,001

Table: Performance of stopping criteria: errors and matvecs (mv) for Test Problem 3 ($m = 10$) for various N_0. Legend: †: out of memory; $-$: does not apply; $*$: does not exist.

$$\text{mv} := \text{matvecs}(m) = \sum_{k=1}^{m} \frac{\text{nnz}(A_k)}{\text{nnz}(A_m)} \cdot \text{its}(k),$$

Generalizations to Mixed FEM

Commutative diagram between Hilbert spaces

$$\begin{tikzcd}
\mathbb{M} \arrow[r, "\mathscr{A}^\star"] \arrow[d, shift left, "\mathscr{M}"] & \mathbb{N}^\star \arrow[d, shift left, "\mathscr{N}^{-1}"] \\
\mathbb{M}^\star \arrow[u, shift left, "\mathscr{M}^{-1}"] & \mathbb{N} \arrow[l, "\mathscr{A}"] \arrow[u, shift left, "\mathscr{N}"]
\end{tikzcd}$$

Problem and theoretical background

Let $\mathbf{M} \in \mathrm{R}^{m\times m}$ and $\mathbf{N} \in \mathrm{R}^{n\times n}$ be symmetric positive definite matrices, and let $\mathbf{A} \in \mathrm{R}^{m\times n}$ $(m \geq n)$ be a full rank matrix. In the following, we will use the following Hilbert spaces

$$\mathbb{M} = \{\mathbf{v} \in \mathrm{R}^m; \|\mathbf{v}\|^2_{\mathbf{M}} = \mathbf{v}^T\mathbf{M}\mathbf{v}\}$$

$$\mathbb{N} = \{\mathbf{q} \in \mathrm{R}^n; \|\mathbf{q}\|^2_{\mathbf{N}} = \mathbf{q}^T\mathbf{N}\mathbf{q}\}$$

and their dual spaces

$$\mathbb{M}^\star = \{\mathbf{w} \in \mathrm{R}^m; \|\mathbf{w}\|^2_{\mathbf{M}^{-1}} = \mathbf{w}^T\mathbf{M}^{-1}\mathbf{w}\}$$

$$\mathbb{N}^\star = \{\mathbf{y} \in \mathrm{R}^n; \|\mathbf{y}\|^2_{\mathbf{N}^{-1}} = \mathbf{y}^T\mathbf{N}^{-1}\mathbf{y}\}.$$

Problem and theoretical background

We will denote by

$$(\mathbf{v}_1, \mathbf{v}_2)_{\mathbf{M}} = \mathbf{v}_1^T \mathbf{M} \mathbf{v}_2, \forall \mathbf{v}_1, \mathbf{v}_2 \in \mathbb{M}$$

and

$$(\mathbf{q}_1, \mathbf{q}_2)_{\mathbf{N}} = \mathbf{q}_1^T \mathbf{N} \mathbf{q}_2, \forall \mathbf{q}_1, \mathbf{q}_2 \in \mathbb{N}$$

the scalar products for $\mathbb{M}$ and $\mathbb{N}$, and by

$$(\mathbf{w}_1, \mathbf{w}_2)_{\mathbf{M}^{-1}} = \mathbf{w}_1^T \mathbf{M}^{-1} \mathbf{w}_2, \forall \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{M}^\star$$

and

$$(\mathbf{y}_1, \mathbf{y}_2)_{\mathbf{N}^{-1}} = \mathbf{y}_1^T \mathbf{N}^{-1} \mathbf{y}_2, \forall \mathbf{y}_1, \mathbf{y}_2 \in \mathbb{N}^\star$$

the respective scalar product for their dual spaces. Finally, we will denote by $\langle \cdot, \cdot \rangle_{\mathbb{M}^\star, \mathbb{M}}$ and by $\langle \cdot, \cdot \rangle_{\mathbb{N}^\star, \mathbb{N}}$, respectively the action of a linear functional on the primal vectors.

Problem and theoretical background

We remark that, using the previous notation, the matrix $\mathbf{A}$ is the representation of a linear operator $\mathscr{A}$ from $\mathbb{N}$ to $\mathbb{M}^\star$. In particular, for each fixed $\mathbf{q} \in \mathbb{N}$ we also have from the Riesz theorem that

$$\langle \mathbf{Aq}, \mathbf{v} \rangle_{\mathbb{M}^\star,\mathbb{M}} = (\mathbf{v}, \mathbf{M}^{-1}\mathbf{Aq})_{\mathbf{M}} = \mathbf{v}^T\mathbf{Aq}, \quad \mathbf{Aq} \in \mathbb{M}^\star \; \forall \mathbf{q} \in \mathbb{N}.$$

Moreover, the matrix $\mathbf{A}^\star$ representing the adjoint operator of $\mathscr{A}$ can be defined as

$$\langle \mathbf{A}^\star\mathbf{g}, \mathbf{f} \rangle_{\mathbb{N}^\star,\mathbb{N}} = (\mathbf{f}, \mathbf{N}^{-1}\mathbf{A}^T\mathbf{g})_{\mathbf{N}} = \mathbf{f}^T\mathbf{A}^T\mathbf{g}, \quad \mathbf{A}^T\mathbf{g} \in \mathbb{N}^\star \; \forall \mathbf{g} \in \mathbb{M},$$

where $\mathbf{A}^\star = \mathbf{N}^{-1}\mathbf{A}^T$.

Problem and theoretical background

We will call the critical points for the functional

$$\sigma = \frac{\mathbf{x}^T \mathbf{A}\mathbf{p}}{\|\mathbf{p}\|_{\mathbf{N}} \, \|\mathbf{x}\|_{\mathbf{M}}} \tag{1}$$

the "*elliptic singular values*" σ_i and the "*elliptic singular vectors*" $\mathbf{p}_i \in \mathbb{N}$ and $\mathbf{x}_i \in \mathbb{M}$, of $\mathbf{A}$.

Mixed FEM

We assume to use RT_0 mixed FEM (Brezzi and Fortin book)

Linear algebra framework

$$\min_{\mathbf{u}}\{\frac{1}{2}\|\mathbf{u}\|^2_{\mathbf{M}} \text{ such that: } \mathbf{A}^T\mathbf{u} = \mathbf{b}, \quad \mathbf{u} \in \mathbb{M},\ \mathbf{b} \in \mathbb{N}^\star\}.$$

The augmented system that gives the optimality conditions for this problem is

$$\begin{bmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix}.$$

Linear algebra framework

Several problems can be reduced to the previous case. The general problem

$$\min_{\mathbf{A}^T\mathbf{w}=\mathbf{r}} \frac{1}{2}\mathbf{w}^T\mathbf{W}\mathbf{w} - \mathbf{g}^T\mathbf{w}$$

where the matrix $\mathbf{W}$ is positive semidefinite and $\ker(\mathbf{W}) \cap \ker(\mathbf{A}^T) = 0$ can be reformulated by choosing $1 \geq \nu \geq 0$ and

$$\left.\begin{array}{l} \mathbf{M} = \mathbf{W} + \nu\mathbf{A}\mathbf{N}^{-1}\mathbf{A}^T \\ \mathbf{u} = \mathbf{w} - \mathbf{M}^{-1}\mathbf{g} \\ \mathbf{b} = \mathbf{r} - \mathbf{A}^T\mathbf{M}^{-1}\mathbf{g}. \end{array}\right\}$$

The non singularity of $\mathbf{M}$ follows from $\ker(\mathbf{W}) \cap \ker(\mathbf{A}^T) = 0$ and the equivalence between the two systems follows from the simple change of variable defined by the second equation.

Linear algebra framework

We point out that the previous transformation is the algebraic version of the preconditioner for the H_{div}-based differential problems described by Arnold, Falk, and Winther, Math. Comp., 1997. In this particular case, the new **M** is the Grammian of the true norm of H_{div} computed on the finite-element test functions used to approximate the continuous problem and in its optimality as a preconditioner is proved by Arnold, Falk, and Winther, Math. Comp., 1997.

Generalized Golub-Kahan Bidiagonalization

$$\begin{cases} \mathbf{AQ} & = & \mathbf{MV}\begin{bmatrix} \mathbf{B} \\ 0 \end{bmatrix} \qquad\qquad \mathbf{V}^T\mathbf{MV} = \mathbf{I}_m \\ \mathbf{A}^T\mathbf{V} & = & \mathbf{NQ}\left[\mathbf{B}^T; 0\right] \qquad\qquad \mathbf{Q}^T\mathbf{NQ} = \mathbf{I}_n \end{cases}$$

where

$$\mathbf{B} = \begin{bmatrix} \alpha_1 & \beta_2 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \beta_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & \alpha_{n-1} & \beta_n \\ 0 & \cdots & 0 & 0 & \alpha_n \end{bmatrix}.$$

The singular values of **B** are linked to the elliptic singular values of **A**:

Generalized Golub-Kahan Bidiagonalization

Algorithm 4.

```
procedure $[\mathbf{U}, \mathbf{V}, \mathbf{B}, \mathbf{u}, \mathbf{p}] =$ G-K_bidiagonalization($\mathbf{A}, \mathbf{M}, \mathbf{N}, \mathbf{b}, maxit$);
  $\beta_1 = \|\mathbf{b}\|_{\mathbf{N}^{-1}}$; $\mathbf{q}_1 = \mathbf{N}^{-1}\mathbf{b}/\beta_1$;
  $\mathbf{w} = \mathbf{M}^{-1}\mathbf{A}\mathbf{q}_1$; $\alpha_1 = \|\mathbf{w}\|_{\mathbf{M}}$; $\mathbf{v}_1 = \mathbf{w}/\alpha_1$;
  $\zeta_1 = \beta_1/\alpha_1$; $\mathbf{d}_1 = \mathbf{q}_1/\alpha_1$; $\mathbf{p}^{(1)} = -\zeta_1\mathbf{d}_1$
  k = 0; convergence = false;
  while convergence = false and $k < maxit$
    $k = k + 1$;
    $\mathbf{g} = \mathbf{N}^{-1}\left(\mathbf{A}^T\mathbf{v}_k - \alpha_k\mathbf{N}\mathbf{q}_k\right)$; $\beta_{k+1} = \|\mathbf{g}\|_{\mathbf{N}}$;
    $\mathbf{q}_{k+1} = \mathbf{g}/\beta_{k+1}$;
    $\mathbf{w} = \mathbf{M}^{-1}\left(\mathbf{A}\mathbf{q}_{k+1} - \beta_{k+1}\mathbf{M}\mathbf{v}_k\right)$; $\alpha_{k+1} = \|\mathbf{w}\|_{\mathbf{M}}$;
    $\mathbf{v}_{k+1} = \mathbf{w}/\alpha_{k+1}$;
    $\zeta_{k+1} = -\frac{\beta_{k+1}}{\alpha_{k+1}}\zeta_k$;
    $\mathbf{d}_{k+1} = \left(\mathbf{q}_{k+1} - \beta_{k+1}\mathbf{d}_k\right)/\alpha_{k+1}$;
    $\mathbf{u}^{(k+1} = \mathbf{u}^{(k)} + \zeta_{k+1}\mathbf{v}_{k+1}$; $\mathbf{p}^{(k+1} = \mathbf{p}^{(k)} - \zeta_{k+1}\mathbf{d}_{k+1}$;
    [ convergence ] = check($\mathbf{z}_k, \ldots$)
  end while;
end procedure.
```

Stopping criteria and error estimates

Let $\mathbf{e}^{(k)} = \mathbf{u} - \mathbf{u}^{(k)}$

$$\|\mathbf{e}^{(k)}\|_{\mathbf{M}}^2 = \sum_{j=k+1}^{n} \zeta_j^2 = \left\| \hat{\mathbf{z}} - \begin{bmatrix} \mathbf{z}_k \\ 0 \end{bmatrix} \right\|_2^2.$$

$$\|\mathbf{p} - \mathbf{p}^{(k)}\|_{\mathbf{N}} = \left\| \mathbf{Q}\mathbf{B}^{-1} \left(\hat{\mathbf{z}} - \begin{bmatrix} \mathbf{z}_k \\ 0 \end{bmatrix} \right) \right\|_{\mathbf{N}} \le \|\mathbf{B}\|_2 \|\mathbf{e}^{(k)}\|_{\mathbf{M}} = \frac{\|\mathbf{e}^{(k)}\|_{\mathbf{M}}}{\sigma_n}.$$

A lower bound estimate

Given a threshold $\tau < 1$ and an integer d, we can estimate $\|\mathbf{e}^{(k)}\|^2_{\mathbf{M}}$ by

$$\xi^2_{k,d} = \sum_{j=k+1}^{k+d+1} \zeta_j^2 < \|\mathbf{e}^{(k)}\|^2_{\mathbf{M}}.$$

An upper bound estimate

It would also be useful to have an upper bound estimator of the error. We can use an approach inspired by the Gauss-Radau quadrature algorithm and similar to the one described by Golub an Meurant (book)

An upper bound estimate

Let $0 < a < \sigma_n$ a lower bound for all the singular values of $\mathbf{B}$. We can then compute the matrix $\hat{\mathbf{T}}_{k+1}$ as

$$\hat{\mathbf{T}}_{k+1} = \left[\begin{array}{cc} \mathbf{T}_k & \alpha_k \beta_k \mathbf{e}_k \\ \alpha_k \beta_k \mathbf{e}_k^T & \omega_{k+1} \end{array} \right],$$

where $\omega_{k+1} = a^2 + \delta_k(a^2)$ and $\delta_k(a^2)$ is the k-entry of the solution of

$$\left(\mathbf{T}_k - a^2 \mathbf{I} \right) \delta(a^2) = \alpha_k^2 \beta_k^2 \mathbf{e}_k.$$

We point out that the matrix $\left(\mathbf{T}_k - a^2 \mathbf{I} \right)$ is positive definite and that $\hat{\mathbf{T}}_{k+1}$ has one eigenvalue equal to a^2.
Analogously to what is done in Golub and Meurant book for the conjugate gradient method, we can recursively compute $\delta(a^2)_k$ and ω_{k+1} by using the Cholesky decomposition.

Mixed finite-element

The aim is to have error bounds merging the approximation error for the mixed finite-element method and the algebraic errors introduced by the generalized G-K bidiagonalization method. Let $\mathbb{H}$ and $\mathbb{P}$ be two Hilbert spaces, and $\mathbb{H}^\star$ and $\mathbb{P}^\star$ the corresponding dual spaces. Let

$$\begin{array}{ll} \mathfrak{a}(u,v) : \mathbb{H} \times \mathbb{H} \to \mathbf{R} & \mathfrak{b}(u,q) : \mathbb{H} \times \mathbb{P} \to \mathbf{R} \\ |\mathfrak{a}(u,v)| \leq \|\mathfrak{a}\| \, \|u\|_{\mathbb{H}} \, \|u\|_{\mathbb{H}} & \forall u \in \mathbb{H}, \forall v \in \mathbb{H} \\ |\mathfrak{b}(u,q)| \leq \|\mathfrak{b}\| \, \|v\|_{\mathbb{H}} \, \|q\|_{\mathbb{P}} & \forall u \in \mathbb{H}, \forall q \in \mathbb{P} \end{array}$$

be continuous bilinear forms with $\|\mathfrak{a}\|$ and $\|\mathfrak{b}\|$ the corresponding norms. Given $f \in \mathbb{H}^\star$ and $g \in \mathbb{P}^\star$, we seek the solutions $u \in \mathbb{H}$ and $p \in \mathbb{P}$ of the system

$$\begin{array}{lll} \mathfrak{a}(u,v) + \mathfrak{b}(v,p) & = \langle f, v \rangle_{\mathbb{H}^\star,\mathbb{H}} & \forall v \in \mathbb{H} \\ \mathfrak{b}(u,q) & = \langle g, q \rangle_{\mathbb{P}^\star,\mathbb{P}} & \forall q \in \mathbb{P}. \end{array}$$

Mixed finite-element

We can introduce the operators $\mathscr{M}$, $\mathscr{A}$ and its adjoint $\mathscr{A}^\star$

$$\begin{array}{lll} \mathscr{M} & : & \mathbb{H} \rightarrow \mathbb{H}^\star, \quad \langle \mathscr{M} u, v\rangle_{\mathbb{H}^\star \times \mathbb{H}} = \mathfrak{a}(u,v) \quad \forall u \in \mathfrak{H}, \forall v \in \mathbb{H} \\ \mathscr{A}^\star & : & \mathbb{H} \rightarrow \mathbb{P}^\star, \quad \langle \mathscr{A}^\star u, q\rangle_{\mathbb{P}^\star \times \mathbb{P}} = \mathfrak{b}(u,q) \quad \forall u \in \mathbb{H}, \forall q \in \mathbb{P} \\ \mathscr{A} & : & \mathbb{P} \rightarrow \mathbb{H}^\star, \quad \langle v, \mathscr{A} p\rangle_{\mathbb{H} \times \mathbb{H}^\star} = \mathfrak{b}(v,p) \quad \forall v \in \mathbb{H}, \forall p \in \mathbb{P} \end{array}$$

and we have

$$\langle \mathscr{A}^\bigstar u, q\rangle_{\mathbb{P}^\star \times \mathbb{P}} = \langle u, \mathscr{A} q\rangle_{\mathbb{H} \times \mathbb{H}^\star} = \mathfrak{b}(u,q).$$

In order to make the following discussion simpler, we assume that $\mathfrak{a}(u,v)$ is symmetric and coercive on $\mathbb{H}$

$$(1) \qquad 0 < \chi_1 \|u\|_{\mathbb{H}} \leq \mathfrak{a}(u,u).$$

However, the coercivity on the kernel of $\mathscr{A}^\star$, $Ker(\mathscr{A}^\star)$ is sufficient.

Mixed finite-element

We will also assume that $\exists \chi_0 > 0$ such that

$$(2) \qquad \sup_{v \in \mathbb{H}} \frac{\mathfrak{b}(v, q)}{\|v\|_{\mathbb{H}}} \geq \chi_0 \|q\|_{\mathbb{P} \setminus Ker(\mathscr{A})} = \chi_0 \left[\inf_{q_0 \in Ker(\mathscr{A})} \|q + q_0\|_{\mathbb{P}} \right].$$

Under the hypotheses (1), (2), and for any $f \in \mathbb{H}^\star$ and $g \in Im(\mathscr{A}^\star)$ then there exists (u, p) solution of the system. Moreover, u is unique and p is definite up to an element of $Ker(\mathscr{A})$.

Mixed finite-element

Let now $\mathbb{H}_h \hookrightarrow \mathbb{H}$ and $\mathbb{P}_h \hookrightarrow \mathbb{P}$ be two finite dimensional subspaces of $\mathbb{H}$ and $\mathbb{P}$. As for the saddle-point problem , we can introduce the operators $\mathscr{A}_h : \mathbb{P}_h \to \mathbb{H}_h^\star$ and $\mathscr{M}_h ; \mathbb{H}_h \to \mathbb{H}_h^\star$. We also assume that

$$(3) \qquad \begin{cases} Ker(\mathscr{A}_h) \subset Ker(\mathscr{A}) \\ \sup_{v_h \in \mathbb{H}_h} \dfrac{\mathfrak{b}(v_h, q_h)}{\|v_h\|_{\mathbb{H}}} \geq \chi_n \|q_h\|_{\mathbb{P}\setminus Ker(\mathscr{A}_h)} \\ \chi_n \geq \chi_0 > 0. \end{cases}$$

Under the hypotheses (1), (2), and (3), we have that $\exists (u_h, p_h) \in \mathbb{H}_h \times \mathbb{P}_h$ solution of

$$\begin{array}{lll} \mathfrak{a}(u_h, v_h) + \mathfrak{b}(v_h, p_h) & = \langle f, v_h \rangle_{\mathbb{H}_h^\star, \mathbb{H}_h} & \forall v_h \in \mathbb{H}_h \\ \mathfrak{b}(u_h, q_h) & = \langle g, q_h \rangle_{\mathbb{P}_h^\star, \mathbb{P}_h} & \forall q_h \in \mathbb{P}_h. \end{array}$$

and

$$\|u - u_h\|_{\mathbb{H}} \quad + \quad \|p - p_h\|_{\mathbb{P}\setminus Ker(A)} \leq \kappa \left(\inf_{v_h \in \mathbb{H}_h} \|u - v_h\|_{\mathbb{H}} + \inf_{q_h \in \mathbb{P}_h} \|p - q_h\|_{\mathbb{P}} \right),$$

where $\kappa = \kappa(\|\mathfrak{a}\|, \|\mathfrak{b}\|, \chi_0, \chi_1)$ is independent of h.

Mixed finite-element

Let $\{\phi_i\}_{i=1,\dots,m}$ be a basis for $\mathbb{H}_h$ and $\left\{\psi_j\right\}_{j=1,\dots,n}$ be a basis for $\mathbb{P}_h$. Then, the matrices $\mathbf{M}$ and $\mathbf{N}$ are the Grammian matrices of the operators $\mathscr{M}$ and $\mathscr{A}$. In order to use the latter theory, we need to weaken the hypothesis, made in the introduction, that $\mathbf{A}$ be full rank. In this case, we have that

- s elliptic singular values will be zero;
- however, the G-K bidiagonalization method will still work and, if $\mathbf{A}\mathbf{q}_1 \neq 0$, it will compute a matrix $\mathbf{B}$ of rank less than or equal to $n - s$.

On the basis of the latter observations, the error $\|\mathbf{e}^{(k)}\|_{\mathbf{M}}$ can be still computed and the upper bounds of the errors computed by G-K hold. Finally, we point out the (**??**) imply that for $h \downarrow 0$ the elliptic singular values of all $\mathbf{A} \in \mathrm{R}^{m_h \times n_h}$ will be bounded with upper and lower bounds independent of h, i.e.

$$\chi_0 \leq \sigma_{n_h} \leq \cdots \leq \sigma_1 \leq \|\mathfrak{a}\|.$$

Mixed finite-element

Theorem

Under (1), (2), and (3), and denoting by $\mathbf{u}^*$ and $\mathbf{p}^*$ the vectors computed at one of the iterates of Algorithm for which $\|\mathbf{e}^{(k)}\|_{\mathbf{M}} < \tau$, we have

$$\begin{aligned}\|u - u^*\|_{\mathbb{H}} \quad + \quad & \|p - p^*\|_{\mathbb{P}\setminus Ker(\mathscr{A})} \leq \\ & \check{\kappa}\left(\inf_{v_h \in \mathbb{H}_h} \|u - v_h\|_{\mathbb{H}} + \inf_{q_h \in \mathbb{P}_h} \|p - q_h\|_{\mathbb{P}} + \tau\right),\end{aligned}$$

where $u^* = \sum_{i=1}^{n_h} \phi_i \mathbf{u}_i^* \in \mathbb{H}_h$, $p^* = \sum_{j=1}^{n_h} \phi_i \mathbf{p}_j^* \in \mathbb{P}_h$ and $\check{\kappa}$ a constant independent of h.

frametitleConclusions?

frametitleConclusions? Can we use the previous framework to build an iAMFEM?

frametitleConclusions? Thank You