Matrices, Graphs, and PDEs:
A journey of unexpected relations

Mario Arioli[1]

[1]BMS Berlin visiting Professor
mario.arioli@gmail.com
http://www3.math.tu-berlin.de/Vorlesungen/SS14/MatricesGraphsPDEs/
Thanks Oliver

Programme

1. Linear Algebra from a variational point of view
2. Short introduction to the Finite Element method (FEM) and adaptive FEM
3. Complex networks

Programme

Linear Algebra from a variational point of view:

- finite dimensional spaces on R^N with a norm based on a positive definite matrix $\mathbf{A}$: a finite dimensional Hilbert spaces theory
- duality and convergence in dual norm
- relations between finite-element approximation matrices and measure of the error in energy
- Golub-Kahan bidiagonalisation method and elliptic singular values

Programme

Short introduction to the Finite Element method (FEM) and adaptive FEM:

- How to use the properties of finite dimensional Hilbert spaces in order to detect where we need to improve the mesh
- Interplay between mesh graphs and matrices
- Fiedler vectors and partitioning of graphs
- Elements of Domain Decomposition techniques
- Adaptive methods and a posteriori measures of the algebraic errors within the Krylov iterative methods

Programme

Complex networks

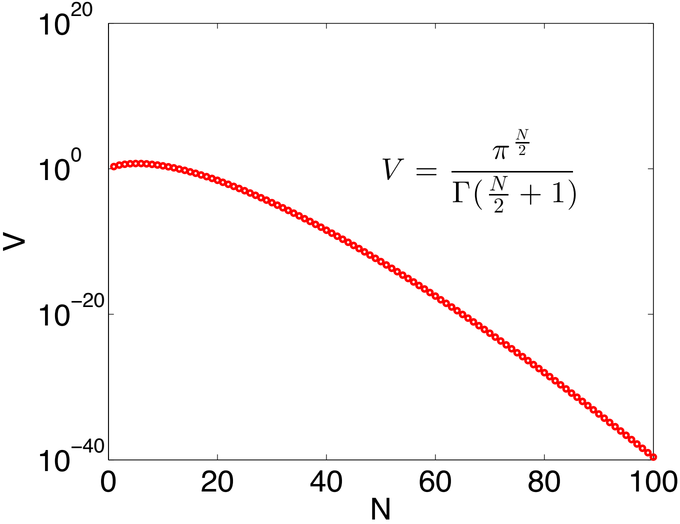
- Elementary introductions to random graphs (Erdos-Renyi, Barabasi-Albert, and Watts- Strogatz random models) and complex graphs2: random models vs real life models
- Embedding of a graph in RN: quantum graphs and 1D-simplex domains
- Solution of systems of parabolic equations on a quantum graph3: Hamiltonians on graphs
- Applications in material science: Dirac's model on graphene

Prologue

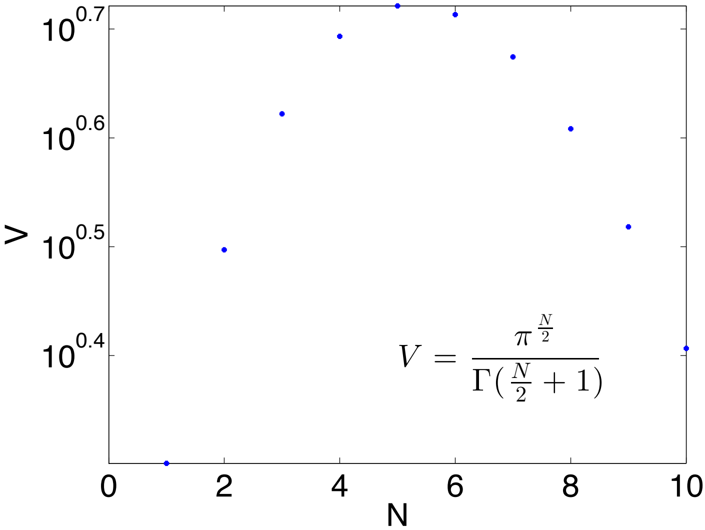
Prologue

When I'm feeling sad
I simply remember my favorite things
And then I don't feel so bad

Les boules


10^{20}
10^{0}
10^{-20}
10^{-40}
V
0
20
40
60
80
100
N
$V = \frac{\pi^{\frac{N}{2}}}{\Gamma(\frac{N}{2}+1)}$

Les boules



$$V = \frac{\pi^{\frac{N}{2}}}{\Gamma(\frac{N}{2}+1)}$$

Les boules

$$V(N) = \frac{\pi^{\frac{N}{2}}}{\Gamma(\frac{N}{2}+1)}$$

N	V
2	π
3	$\frac{4}{3}\pi$
4	$\frac{\pi^2}{2}$
5	$\frac{8\pi^2}{15}$
6	$\frac{\pi^3}{6}$
10	$\frac{\pi^5}{150}$

Les boules

$$V(N) = \frac{\pi^{\frac{N}{2}}}{\Gamma(\frac{N}{2}+1)}$$

IS the Volume of the N-dimensional Sphere and

$$\lim_{N\to\infty} V(N) \to 0$$

Les boules

$$\mathbf{v} = \frac{1}{\sqrt{N}} * \mathbf{e}_N = \left.\begin{bmatrix} \frac{1}{\sqrt{N}} \\ \frac{1}{\sqrt{N}} \\ \vdots \\ \frac{1}{\sqrt{N}} \end{bmatrix}\right\} N$$

$$||\mathbf{v}||_2 = 1$$

BUT

$$\lim_{N \to \infty} \mathbf{v} = \mathbf{0}$$

Les boules

FUNCTIONAL ANALYSIS IS LINEAR ALGEBRA COPING WITH STRANGE BALLS

Some friends

$$\mathbf{G} = \begin{bmatrix}
1 & -1 & & & & & & & & & \\
& 1 & -1 & & & & & & & & \\
& & 1 & -1 & & & & & & & \\
& & & 1 & -1 & & & & & & \\
& & & & 1 & -1 & & & & & \\
& & & & & 1 & -1 & & & & \\
& & & & & & 1 & -1 & & & \\
& & & & & & & 1 & -1 & & \\
& & & & & & & & 1 & -1 & \\
& & & & & & & & & 1 & -1
\end{bmatrix}$$

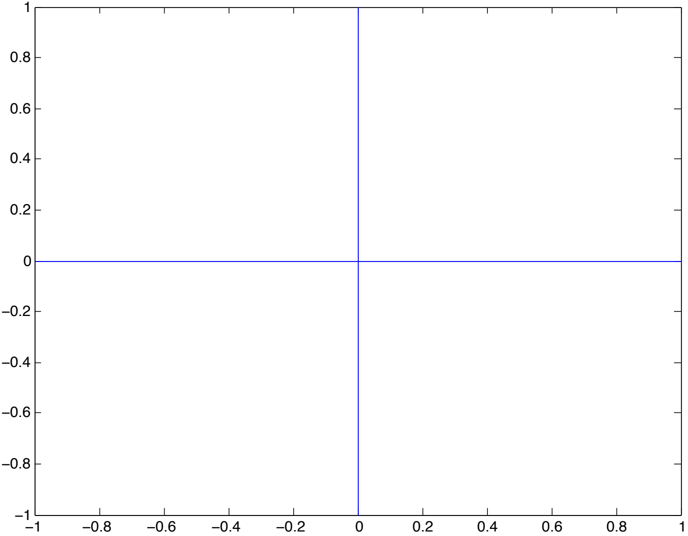
$\frac{d}{dt}$ on the interval $[0, 1]$

Some friends

$$\mathbf{G}^T\mathbf{G} = \begin{bmatrix} 1 & -1 & & & & & & & & & \\ -1 & 2 & -1 & & & & & & & & \\ & -1 & 2 & -1 & & & & & & & \\ & & -1 & 2 & -1 & & & & & & \\ & & & -1 & 2 & -1 & & & & & \\ & & & & -1 & 2 & -1 & & & & \\ & & & & & -1 & 2 & -1 & & & \\ & & & & & & -1 & 2 & -1 & & \\ & & & & & & & -1 & 2 & -1 & \\ & & & & & & & & -1 & 2 & -1 \\ & & & & & & & & & -1 & 1 \end{bmatrix}$$

$\frac{d^2}{(dt)^2}$ Laplacian with Neumann conditions on the interval $[0, 1]$

Some friends



Some friends

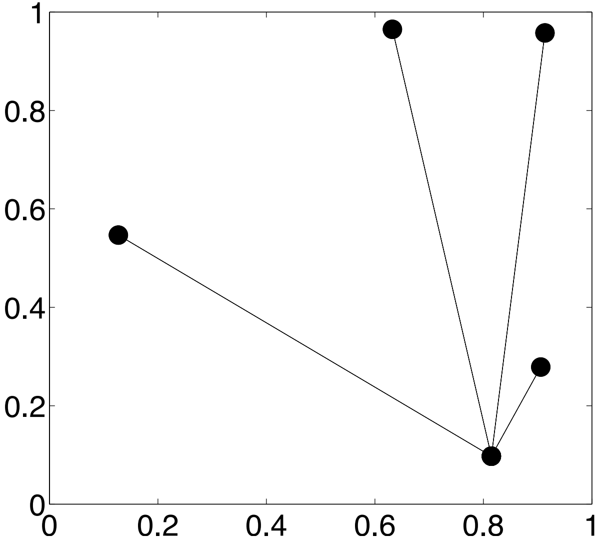
$$\mathbf{G} = \begin{bmatrix} 1 & -1 & & & \\ 1 & & -1 & & \\ 1 & & & -1 & \\ 1 & & & & -1 \end{bmatrix}$$

$\mathbf{G}$ is the INCIDENCE MATRIX of the graph and the grad operator.
$\mathbf{G}^T\mathbf{G}$ is the LAPLACIAN on the graph.
$\mathbf{I} - \mathbf{G}^T\mathbf{G}$ is the ADJACENCY matrix of the graph.

Some friends

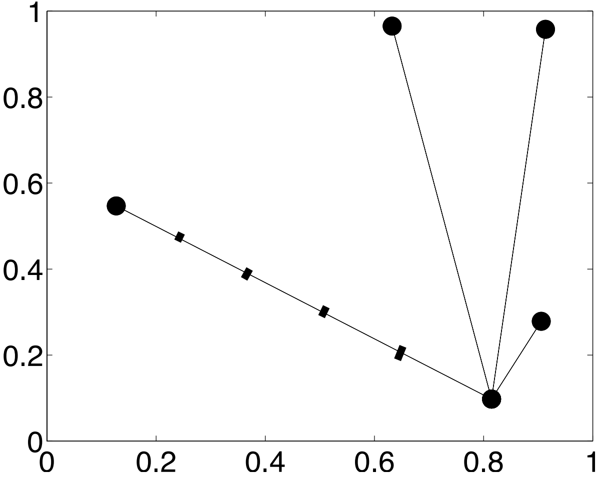
We do not need a regular graph. On each edge, we have a map from the $\mathbf{R}^N$ space where it lives to the segment $[0, 1]$ and we can solve on each edge the local operator!!

Some friends

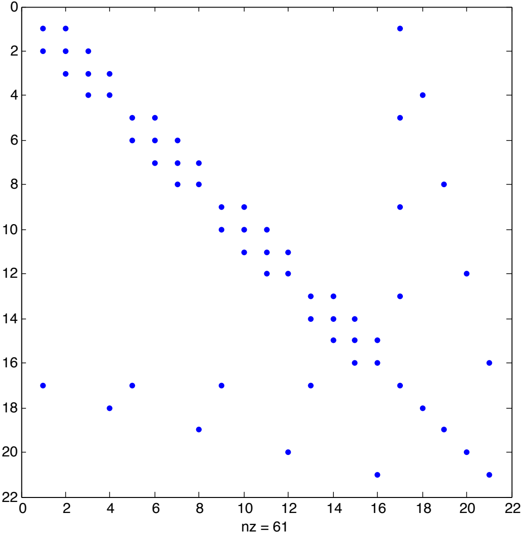


Some friends

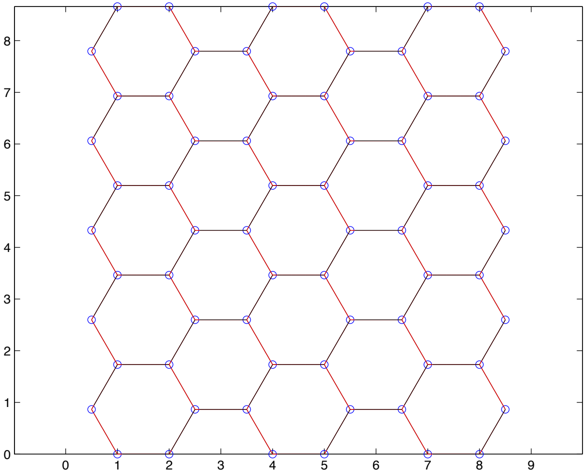
We can insert points on each edge



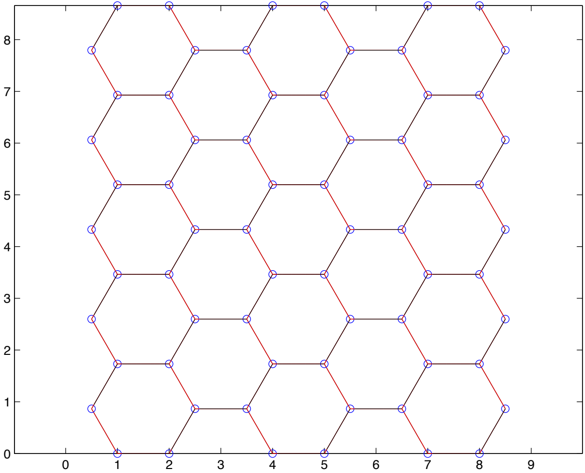
Some friends



Some friends



Some friends



GRAPHENE

TRUTH may be misleading

In a finite dimensional space all norms are equivalent i.e.

TRUTH may be misleading

In a finite dimensional space all norms are equivalent i.e.

$$c(N)||v||_1 \leq ||v||_2 \leq C(N)||v||_1$$

TRUTH may be misleading

In a finite dimensional space all norms are equivalent i.e.

$$c(N)||v||_1 \le ||v||_2 \le C(N)||v||_1$$

Identify the norms for which we have

$$c||v||_1 \le ||v||_2 \le C||v||_1$$

TRUTH may be misleading

In a finite dimensional space all norms are equivalent i.e.

$$c(N)||v||_1 \leq ||v||_2 \leq C(N)||v||_1$$

Identify the norms for which we have

$$c||v||_1 \leq ||v||_2 \leq C||v||_1 \quad \text{i.e.} \quad ||\cdot||_1 \sim ||\cdot||_2$$

Lecture 1

Finite dimensional Hilbert spaces and $\mathbb{R}^N$

- $(\cdot,\cdot) : \mathfrak{H} \times \mathfrak{H} \to \mathbb{R}$ scalar product and
  $\|u\|_{\mathfrak{H}} = \sqrt{(u,u)} \qquad \forall u \in \mathfrak{H}$ norm.

Finite dimensional Hilbert spaces and $\mathbb{R}^N$

- $(\cdot,\cdot) : \mathfrak{H} \times \mathfrak{H} \to \mathbb{R}$ scalar product and
  $\|u\|_{\mathfrak{H}} = \sqrt{(u,u)} \qquad \forall u \in \mathfrak{H}$ norm.
- $\exists\{\psi_i\}_{i=1,\ldots,N}$ a basis for $\mathfrak{H}$
  $\forall u \in \mathfrak{H} \qquad u = \sum_{i=1}^{N} u_i\psi_i \qquad u_i \in \mathbb{R} \quad i = 1,\ldots,N$

Finite dimensional Hilbert spaces and R^N

- $(\cdot,\cdot) : \mathfrak{H} \times \mathfrak{H} \to \mathrm{R}$ scalar product and
  $\|u\|_{\mathfrak{H}} = \sqrt{(u,u)} \qquad \forall u \in \mathfrak{H}$ norm.
- $\exists \{\psi_i\}_{i=1,\dots,N}$ a basis for $\mathfrak{H}$
  $\forall u \in \mathfrak{H} \qquad u = \sum_{i=1}^{N} u_i \psi_i \qquad u_i \in \mathrm{R} \quad i = 1, \dots, N$
- Representation of scalar product in R^N.
  Let $u = \sum_{i=1}^{N} u_i \psi_i$ and $v = \sum_{i=1}^{N} v_i \psi_i$.
  Then

$$(u,v) = \sum_{i=1}^{N} \sum_{j=1}^{N} u_i v_j (\psi_i, \psi_j) = \mathbf{v}^T \mathbf{H} \mathbf{u}$$

where $\mathbf{H}_{ij} = \mathbf{H}_{ji} = (\psi_i, \psi_j)$ and $\mathbf{u}, \mathbf{v} \in \mathrm{R}^N$.
Moreover, $\mathbf{u}^T \mathbf{H} \mathbf{u} > 0 \qquad \text{iff} \qquad \mathbf{u} \neq 0$ and, thus $\mathbf{H}$ SPD.

Dual space $\mathfrak{H}^\star$

- $f \in \mathfrak{H}^\star : \mathfrak{H} \to \mathrm{R}$ (functional);

Dual space $\mathfrak{H}^\star$

- $f \in \mathfrak{H}^\star : \mathfrak{H} \to \mathrm{R}$ (functional);
- $f(\alpha u + \beta v) = \alpha f(u) + \beta f(v) \qquad \forall u, v \in \mathfrak{H}$

Dual space $\mathfrak{H}^\star$

- $f \in \mathfrak{H}^\star : \mathfrak{H} \to \mathrm{R}$ (functional);
- $f(\alpha u + \beta v) = \alpha f(u) + \beta f(v) \qquad \forall u, v \in \mathfrak{H}$
- $\mathfrak{H}^\star$ is the space of the linear functionals on $\mathfrak{H}$

$$\|f\|_{\mathfrak{H}}^\star = \sup_{u \neq 0} \frac{f(u)}{\|u\|_{\mathfrak{H}}}$$

Dual space $\mathfrak{H}^\star$

- $f \in \mathfrak{H}^\star : \mathfrak{H} \to \mathrm{R}$ (functional);
- $f(\alpha u + \beta v) = \alpha f(u) + \beta f(v) \qquad \forall u, v \in \mathfrak{H}$
- $\mathfrak{H}^\star$ is the space of the linear functionals on $\mathfrak{H}$

$$\|f\|_{\mathfrak{H}}^{\star} = \sup_{u \neq 0} \frac{f(u)}{\|u\|_{\mathfrak{H}}}$$

- If $\mathfrak{H}$ finite dimensional and $u = \sum_{i=1}^{N} u_i \psi_i$, then $f(u) = \sum_{i=1}^{N} u_i f(\psi_i) = \mathbf{f}^T \mathbf{u}$

Dual space $\mathfrak{H}^\star$

- $f \in \mathfrak{H}^\star : \mathfrak{H} \to \mathbf{R}$ (functional);
- $f(\alpha u + \beta v) = \alpha f(u) + \beta f(v) \qquad \forall u, v \in \mathfrak{H}$
- $\mathfrak{H}^\star$ is the space of the linear functionals on $\mathfrak{H}$

$$\|f\|^\star_{\mathfrak{H}} = \sup_{u \neq 0} \frac{f(u)}{\|u\|_{\mathfrak{H}}}$$

- If $\mathfrak{H}$ finite dimensional and $u = \sum_{i=1}^N u_i \psi_i$, then $f(u) = \sum_{i=1}^N u_i f(\psi_i) = \mathbf{f}^T \mathbf{u}$
- Dual vector
  Let $u \in \mathfrak{H},\ u \neq 0$, then $\exists f_u \in \mathfrak{H}^\star$ such that

$$f_u(u) = \|u\|_{\mathfrak{H}}$$

(Hahn-Banach).

Dual space $\mathfrak{H}^\star$

- Let $\mathfrak{H}$ be a Hilbert finite dimensional space and $\mathbf{H}$ the real $N \times N$ matrix identifying the scalar product.

Dual space $\mathfrak{H}^\star$

- Let $\mathfrak{H}$ be a Hilbert finite dimensional space and $\mathbf{H}$ the real $N \times N$ matrix identifying the scalar product.
- $$f_u(u) = \mathbf{f}^T\mathbf{u} \quad = \quad (\mathbf{u}^T\mathbf{H}\mathbf{u})^{1/2}$$

  The dual vector of $\mathbf{u}$ has the following representation:

Dual space $\mathfrak{H}^\star$

- Let $\mathfrak{H}$ be a Hilbert finite dimensional space and $\mathbf{H}$ the real $N \times N$ matrix identifying the scalar product.
- $$f_u(u) = \mathbf{f}^T\mathbf{u} \quad = \quad (\mathbf{u}^T\mathbf{H}\mathbf{u})^{1/2}$$

  The dual vector of $\mathbf{u}$ has the following representation:

  $$\mathbf{f} = \frac{\mathbf{H}\mathbf{u}}{\|\mathbf{u}\|_{\mathbf{H}}}$$

  and

  $$\|f_u\|^2_{\mathfrak{H}^\star} = \mathbf{u}^T\mathbf{H}\mathbf{u} = \mathbf{f}^T\mathbf{H}^{-1}\mathbf{f}$$

Dual space basis

- The general definitions of a dual basis for $\mathfrak{H}$ is

$$\phi_j(\psi_i) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Dual space basis

- The general definitions of a dual basis for $\mathfrak{H}$ is

$$\phi_j(\psi_i) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

- The ϕ_i are linearly independent:

$$\sum_{i=1}^{N} \beta_i \phi_i(u) = 0 \ \ \forall u \in \mathfrak{H} \Longrightarrow \sum_{i=1}^{N} \beta_i \phi_i(\psi_i) = 0 \Longrightarrow \beta_i = 0.$$

Dual space basis

- The general definitions of a dual basis for $\mathfrak{H}$ is

$$\phi_j(\psi_i) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

- The ϕ_i are linearly independent:

$$\sum_{i=1}^{N} \beta_i \phi_i(u) = 0 \ \ \forall u \in \mathfrak{H} \Longrightarrow \sum_{i=1}^{N} \beta_i \phi_i(\psi_i) = 0 \Longrightarrow \beta_i = 0.$$

- $f(\psi_i) = \gamma_i$ and $f(u) = f(\sum_{i=1}^{N} u_i \psi_i) = \sum_{i=1}^{N} \gamma_i u_i$

$$\phi_i(u) = \phi(\sum_{i=1}^{N} u_i \psi_i) = u_i \Longrightarrow f = \sum_{i=1}^{N} \alpha_i \phi_i$$

Linear operator

- $\mathscr{A}:\mathfrak{H}\longrightarrow\mathfrak{V}$ where $\mathfrak{H}$ and $\mathfrak{V}$ finite dimensional Hilbert spaces. **H** and **V** are the SPD matrices of the scalar products

Linear operator

- $\mathscr{A} : \mathfrak{H} \longrightarrow \mathfrak{V}$ where $\mathfrak{H}$ and $\mathfrak{V}$ finite dimensional Hilbert spaces. **H** and **V** are the SPD matrices of the scalar products
- $$\|\mathscr{A}\|_{\mathfrak{H},\mathfrak{V}} = \max_{u\neq 0} \frac{\|\mathscr{A}u\|_{\mathfrak{V}}}{\|u\|_{\mathfrak{H}}} = \|\mathbf{V}^{1/2}\mathbf{A}\mathbf{H}^{-1/2}\|_2$$

Linear operator

- $\mathscr{A} : \mathfrak{H} \longrightarrow \mathfrak{V}$ where $\mathfrak{H}$ and $\mathfrak{V}$ finite dimensional Hilbert spaces. **H** and **V** are the SPD matrices of the scalar products
-

$$\|\mathscr{A}\|_{\mathfrak{H},\mathfrak{V}} = \max_{u \neq 0} \frac{\|\mathscr{A} u\|_{\mathfrak{V}}}{\|u\|_{\mathfrak{H}}} = \|\mathbf{V}^{1/2}\mathbf{A}\mathbf{H}^{-1/2}\|_2$$

- The result follows from the generalized eigenvalue problem in R^N

$$\mathbf{A}^T\mathbf{V}\mathbf{A}\mathbf{u} = \lambda\mathbf{H}\mathbf{u}$$

Linear operator

- $\mathscr{A} : \mathfrak{H} \longrightarrow \mathfrak{V}$ where $\mathfrak{H}$ and $\mathfrak{V}$ finite dimensional Hilbert spaces. **H** and **V** are the SPD matrices of the scalar products
- $$\|\mathscr{A}\|_{\mathfrak{H},\mathfrak{V}} = \max_{u \neq 0} \frac{\|\mathscr{A}u\|_{\mathfrak{V}}}{\|u\|_{\mathfrak{H}}} = \|\mathbf{V}^{1/2}\mathbf{A}\mathbf{H}^{-1/2}\|_2$$
- The result follows from the generalized eigenvalue problem in R^N
$$\mathbf{A}^T\mathbf{V}\mathbf{A}\mathbf{u} = \lambda\mathbf{H}\mathbf{u}$$
- $$\kappa_{\mathbf{H}}(\mathbf{M}) = \|\mathbf{M}\|_{\mathbf{H},\mathbf{H}^{-1}}\|\mathbf{M}^{-1}\|_{\mathbf{H}^{-1},\mathbf{H}}.$$

Linear operator

- $\mathscr{A} : \mathfrak{H} \longrightarrow \mathfrak{V}$ where $\mathfrak{H}$ and $\mathfrak{V}$ finite dimensional Hilbert spaces. **H** and **V** are the SPD matrices of the scalar products
- $$\|\mathscr{A}\|_{\mathfrak{H},\mathfrak{V}} = \max_{u \neq 0} \frac{\|\mathscr{A}u\|_{\mathfrak{V}}}{\|u\|_{\mathfrak{H}}} = \|\mathbf{V}^{1/2}\mathbf{A}\mathbf{H}^{-1/2}\|_2$$
- The result follows from the generalized eigenvalue problem in R^N
$$\mathbf{A}^T\mathbf{V}\mathbf{A}\mathbf{u} = \lambda\mathbf{H}\mathbf{u}$$
- $$\kappa_{\mathbf{H}}(\mathbf{M}) = \|\mathbf{M}\|_{\mathbf{H},\mathbf{H}^{-1}}\|\mathbf{M}^{-1}\|_{\mathbf{H}^{-1},\mathbf{H}}.$$
The interesting case is $\kappa_{\mathbf{H}}(\mathbf{M})$ independent of N

Hilbert Space Setting: duality and adjoint

Given $z \in \mathfrak{H}^\star$, we have

$$\langle z, u\rangle_{\mathfrak{H}^\star,\mathfrak{H}} = \mathbf{z}^T\mathbf{u} = \mathbf{z}^T\mathbf{H}^{-1}\mathbf{H}\mathbf{u} = (\mathbf{u}, \mathbf{H}^{-1}\mathbf{z})_{\mathbf{H}},$$

$\mathbf{w} = \mathbf{H}^{-1}\mathbf{z}$ Riesz vector corresponding to $w = \sum_j w_j\phi_j \in \mathfrak{H}$.

Hilbert Space Setting: duality and adjoint

Given $z \in \mathfrak{H}^\star$, we have

$$\langle z, u\rangle_{\mathfrak{H}^\star,\mathfrak{H}} = \mathbf{z}^T\mathbf{u} = \mathbf{z}^T\mathbf{H}^{-1}\mathbf{H}\mathbf{u} = (\mathbf{u}, \mathbf{H}^{-1}\mathbf{z})_{\mathbf{H}},$$

$\mathbf{w} = \mathbf{H}^{-1}\mathbf{z}$ Riesz vector corresponding to $w = \sum_j w_j\phi_j \in \mathfrak{H}$.

Let $\mathscr{C} : \mathfrak{H} \mapsto \mathfrak{F}$
$\mathscr{C}^\star : \mathfrak{F}^\star \mapsto \mathfrak{H}^\star$ (adjoint operator)

$$\langle \mathscr{C}^\star v, u\rangle_{\mathfrak{H}^\star,\mathfrak{H}} \triangleq \langle v, \mathscr{C}u\rangle_{\mathfrak{F}^\star,\mathfrak{F}} \quad \forall v \in \mathfrak{F}^\star, u \in \mathfrak{H}.$$

Hilbert Space Setting: duality and adjoint

Given $z \in \mathfrak{H}^\star$, we have

$$\langle z, u\rangle_{\mathfrak{H}^\star,\mathfrak{H}} = \mathbf{z}^T\mathbf{u} = \mathbf{z}^T\mathbf{H}^{-1}\mathbf{H}\mathbf{u} = (\mathbf{u}, \mathbf{H}^{-1}\mathbf{z})_{\mathbf{H}},$$

$\mathbf{w} = \mathbf{H}^{-1}\mathbf{z}$ Riesz vector corresponding to $w = \sum_j w_j\phi_j \in \mathfrak{H}$.

Let $\mathscr{C} : \mathfrak{H} \mapsto \mathfrak{F}$
$\mathscr{C}^\star : \mathfrak{F}^\star \mapsto \mathfrak{H}^\star$ (adjoint operator)

$$\langle \mathscr{C}^\star v, u\rangle_{\mathfrak{H}^\star,\mathfrak{H}} \triangleq \langle v, \mathscr{C}u\rangle_{\mathfrak{F}^\star,\mathfrak{F}} \quad \forall v \in \mathfrak{F}^\star, u \in \mathfrak{H}.$$

Therefore, we have

$$\langle \mathscr{C}^\star v, u\rangle_{\mathfrak{H}^\star,\mathfrak{H}} = (\mathbf{C}\mathbf{u}, \mathbf{F}^{-1}\mathbf{v})_{\mathbf{F}} = \mathbf{u}^T\mathbf{C}^T\mathbf{v}.$$

Hilbert Space Setting: normal equations

If we assume that $\mathfrak{F} = \mathfrak{H}^\star$ then we have that the "normal equations operator" in the Hilbert space is an operator such that

$$\mathscr{C}^\star \circ \mathscr{H}^{-1} \circ \mathscr{C} : \mathfrak{H} \mapsto \mathfrak{H}^\star,$$

and it is represented by the matrix

$$\mathbf{C}^T \mathbf{H}^{-1} \mathbf{C}.$$

Hilbert Space Setting: normal equations

If we assume that $\mathfrak{F} = \mathfrak{H}^\star$ then we have that the "normal equations operator" in the Hilbert space is an operator such that

$$\mathscr{C}^\star \circ \mathscr{H}^{-1} \circ \mathscr{C} : \mathfrak{H} \mapsto \mathfrak{H}^\star,$$

and it is represented by the matrix

$$\mathbf{C}^T \mathbf{H}^{-1} \mathbf{C}.$$

If $\mathbf{C}^T = \mathbf{C}$ then the corresponding operator $\mathscr{C}$ is self-adjoint. Moreover, we have that the operator

$$\mathscr{H}^{-1} \circ \mathscr{C} : \mathfrak{H} \mapsto \mathfrak{H}$$

maps $\mathfrak{H}$ into itself.

$$\left(\mathscr{H}^{-1} \circ \mathscr{C}\right)^i \triangleq (\mathbf{H}^{-1}\mathbf{C})^i.$$

Linear operators

Let us consider now the Hilbert spaces

$$\mathfrak{M} := (\mathsf{R}^n, \|\cdot\|_{\mathbf{M}}), \qquad \mathfrak{N} := (\mathsf{R}^m, \|\cdot\|_{\mathbf{N}}),$$

and their dual spaces

$$\mathfrak{M}^\star := (\mathsf{R}^n, \|\cdot\|_{\mathbf{M}^{-1}}), \qquad \mathfrak{N}^\star := (\mathsf{R}^m, \|\cdot\|_{\mathbf{N}^{-1}}),$$

Linear operators

Let us consider now the Hilbert spaces

$$\mathfrak{M} := (\mathsf{R}^n, \|\cdot\|_{\mathbf{M}}), \qquad \mathfrak{N} := (\mathsf{R}^m, \|\cdot\|_{\mathbf{N}}),$$

and their dual spaces

$$\mathfrak{M}^\star := (\mathsf{R}^n, \|\cdot\|_{\mathbf{M}^{-1}}), \qquad \mathfrak{N}^\star := (\mathsf{R}^m, \|\cdot\|_{\mathbf{N}^{-1}}),$$

$$\mathscr{A} : \mathfrak{N} \to \mathfrak{M}^\star$$

$$\langle \mathscr{A} y, u \rangle_{\mathfrak{M}^\star, \mathfrak{M}} \triangleq (\mathbf{u}, \mathbf{M}^{-1}\mathbf{A}\mathbf{y})_{\mathbf{M}} = \mathbf{u}^T\mathbf{A}\mathbf{y}, \quad y \in \mathfrak{N}, \forall u \in \mathfrak{M},$$

Linear operators

Let us consider now the Hilbert spaces

$$\mathfrak{M} := (\mathsf{R}^n, \|\cdot\|_{\mathbf{M}}), \qquad \mathfrak{N} := (\mathsf{R}^m, \|\cdot\|_{\mathbf{N}}),$$

and their dual spaces

$$\mathfrak{M}^\star := (\mathsf{R}^n, \|\cdot\|_{\mathbf{M}^{-1}}), \qquad \mathfrak{N}^\star := (\mathsf{R}^m, \|\cdot\|_{\mathbf{N}^{-1}}),$$

$$\mathscr{A} : \mathfrak{N} \to \mathfrak{M}^\star$$

$$\langle \mathscr{A} y, u \rangle_{\mathfrak{M}^\star,\mathfrak{M}} \triangleq (\mathbf{u}, \mathbf{M}^{-1}\mathbf{A}\mathbf{y})_{\mathbf{M}} = \mathbf{u}^T\mathbf{A}\mathbf{y}, \quad y \in \mathfrak{N}, \forall u \in \mathfrak{M},$$

$$\langle \mathscr{A}^\star u, y \rangle_{\mathfrak{N}^\star,\mathfrak{N}} := (\mathbf{y}, \mathbf{N}^{-1}\mathbf{A}^T\mathbf{u})_{\mathbf{N}} = \mathbf{y}^T\mathbf{A}^T\mathbf{u}, \quad u \in \mathfrak{M}, \forall y \in \mathfrak{N},$$

Linear operators

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

Linear operators

$$\mathbf{C} = \begin{bmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & -\mathbf{N} \end{bmatrix}$$

$$\mathscr{C} : \mathfrak{M} \times \mathfrak{N} \mapsto \mathfrak{M}^\star \times \mathfrak{N}^\star.$$

The scalar product in $\mathfrak{M} \times \mathfrak{N}$ is represented by the matrix

$$\mathbf{H} = \begin{bmatrix} \mathbf{M} & \\ & \mathbf{N} \end{bmatrix}.$$

Lecture 2

Linear systems: variational framework

- Find $\mathbf{u} \in \mathfrak{H}$ such that for all $\mathbf{v} \in \mathfrak{H}$

$$a(\mathbf{u}, \mathbf{v}) = L(\mathbf{v}) \qquad (L(\cdot) \in \mathfrak{H}^\star \text{ dual space of } \mathfrak{H})$$

Linear systems: variational framework

- Find $\mathbf{u} \in \mathfrak{H}$ such that for all $\mathbf{v} \in \mathfrak{H}$

$$a(\mathbf{u}, \mathbf{v}) = L(\mathbf{v}) \qquad (L(\cdot) \in \mathfrak{H}^\star \text{ dual space of } \mathfrak{H})$$

- Existence and uniqueness: $\forall \mathbf{v}, \mathbf{w} \in \mathfrak{H}$

$$\begin{array}{rcl} a(\mathbf{w}, \mathbf{v}) & \leq & C_1 \|\mathbf{w}\|_{\mathfrak{H}} \|\mathbf{v}\|_{\mathfrak{H}} \\ \sup\limits_{\mathbf{w} \in \mathfrak{H} \setminus \{0\}} \dfrac{a(\mathbf{w}, \mathbf{v})}{\|\mathbf{w}\|_{\mathfrak{H}}} & \geq & C_2 \|\mathbf{v}\|_{\mathfrak{H}} \end{array}$$

Linear systems: variational framework

- Find $\mathbf{u} \in \mathfrak{H}$ such that for all $\mathbf{v} \in \mathfrak{H}$

$$a(\mathbf{u}, \mathbf{v}) = L(\mathbf{v}) \qquad (L(\cdot) \in \mathfrak{H}^{\star} \ \text{dual space of } \mathfrak{H})$$

- Existence and uniqueness: $\forall \mathbf{v}, \mathbf{w} \in \mathfrak{H}$

$$\begin{aligned} a(\mathbf{w}, \mathbf{v}) &\leq C_1 \|\mathbf{w}\|_{\mathfrak{H}} \|\mathbf{v}\|_{\mathfrak{H}} \\ \sup_{\mathbf{w} \in \mathfrak{H} \setminus \{0\}} \frac{a(\mathbf{w}, \mathbf{v})}{\|\mathbf{w}\|_{\mathfrak{H}}} &\geq C_2 \|\mathbf{v}\|_{\mathfrak{H}} \end{aligned}$$

- $\mathfrak{H} = (\mathrm{R}^N, ||\cdot||_{\mathbf{H}})$ and $\mathfrak{H}^{\star} = (\mathrm{R}^N, ||\cdot||_{\mathbf{H}^{-1}})$
  H SPD

Finite dimensional Banach spaces and $\mathbb{R}^N$

- Other norms are possible on $\mathbf{R}^N$:

Finite dimensional Banach spaces and $\mathbb{R}^N$

- Other norms are possible on $\mathbf{R}^N$:
  - $\|\mathbf{u}\|_p = (\sum_{i=1}^N (u_i)^p)^{1/2}$ with $1 < p < \infty$
  - $\|\mathbf{u}\|_1 = (\sum_{i=1}^N |u_i|)$
  - $\|\mathbf{u}\|_\infty = \max_i |u_i|$

Finite dimensional Banach spaces and R^N

- Other norms are possible on $\mathbf{R}^N$:
  - $\|\mathbf{u}\|_p = (\sum_{i=1}^N (u_i)^p)^{1/2}$ with $1 < p < \infty$
  - $\|\mathbf{u}\|_1 = (\sum_{i=1}^N |u_i|)$
  - $\|\mathbf{u}\|_\infty = \max_i |u_i|$
- Hyper-norms on $\mathbf{R}^N$ of order k.

$$\|\cdot\|_{\vec{k}}: \qquad \mathbf{R}^N \to \mathbf{R}^k$$

$$\begin{array}{lll}
I & \forall \lambda \in \mathbf{R} & \|\lambda \mathbf{u}\|_{\vec{k}} = |\lambda| \|\mathbf{u}\|_{\vec{k}} \\
II & \forall \mathbf{u}, \mathbf{v} \in \mathbf{R}^N & \|\mathbf{u}+\mathbf{v}\|_{\vec{k}} \leq \|\mathbf{u}\|_{\vec{k}} + \|\mathbf{v}\|_{\vec{k}} \text{ component-wise} \\
III & & \|\mathbf{u}\|_{\vec{k}} = 0_k \Rightarrow u = 0_N
\end{array}$$

Linear operator hyper-norm

- Let $\|\mathbf{u}\|_{\vec{k}}$ and $\|\mathbf{v}\|_{\vec{p}}$ two hyper-norms on R^n and $\mathscr{A}$ a linear operator between $(\mathrm{R}^n, \|\cdot\|_{\vec{k}})$ and $(\mathrm{R}^n, \|\cdot\|_{\vec{p}})$

Linear operator hyper-norm

- Let $\|\mathbf{u}\|_{\vec{k}}$ and $\|\mathbf{v}\|_{\vec{p}}$ two hyper-norms on R^n and $\mathscr{A}$ a linear operator between $(\mathrm{R}^n, \|\cdot\|_{\vec{k}})$ and $(\mathrm{R}^n, \|\cdot\|_{\vec{p}})$
- The norm is defined as

$$\begin{aligned}
\|\mathscr{A}\|_{\vec{k},\vec{p}} &= \mathbf{M} \in \mathrm{R}^{k\times p} \\
\mathbf{M} &= \begin{bmatrix} \|\mathbf{A}_{11}\| & \dots & \|\mathbf{A}_{1k}\| \\ \vdots & \dots & \vdots \\ \|\mathbf{A}_{p1}\| & \dots & \|\mathbf{A}_{pk}\| \end{bmatrix}
\end{aligned}$$

Linear operator hyper-norm

- Let $\|\mathbf{u}\|_{\vec{k}}$ and $\|\mathbf{v}\|_{\vec{p}}$ two hyper-norms on R^n and $\mathscr{A}$ a linear operator between $(\mathrm{R}^n, \|\cdot\|_{\vec{k}})$ and $(\mathrm{R}^n, \|\cdot\|_{\vec{p}})$
- The norm is defined as

$$\begin{aligned} \|\mathscr{A}\|_{\vec{k},\vec{p}} &= \mathbf{M} \in \mathrm{R}^{k\times p} \\ \mathbf{M} &= \begin{bmatrix} \|\mathbf{A}_{11}\| & \dots & \|\mathbf{A}_{1k}\| \\ \vdots & \dots & \vdots \\ \|\mathbf{A}_{p1}\| & \dots & \|\mathbf{A}_{pk}\| \end{bmatrix} \end{aligned}$$

-

$$\mathrm{R}^n = \bigoplus_{j=1}^{p} \mathfrak{W}_j = \bigoplus_{i=1}^{k} \mathfrak{V}_i \qquad \mathfrak{W}_i \cap \mathfrak{W}_j = \{0\} \qquad \mathfrak{V}_i \cap \mathfrak{V}_j = \{0\}$$

Rigal-Gaches (1967) theorem

$$\left.\begin{array}{l} \exists \Delta \mathbf{A}, \exists \delta \mathbf{b} \text{ such that:}\\ (\mathbf{A}+\Delta\mathbf{A})\tilde{\mathbf{u}} = (\mathbf{b}+\delta\mathbf{b}) \text{ and}\\ \|\Delta\mathbf{A}\|_{\vec{k},\vec{p}} \le \mathbf{S} \in \mathrm{R}^{k\times p}, \|\delta\mathbf{b}\|_{\vec{k}} \le \mathbf{t}\in\mathrm{R}^k \end{array}\right\} \Leftrightarrow \left\{\begin{array}{l} \|\mathbf{r}\|_{\vec{k}} \le \mathbf{S}\|\tilde{\mathbf{u}}\|_{\vec{p}} + \mathbf{t}\\ \text{where } \mathbf{r} \text{ is defined by}\\ \mathbf{r} = \mathbf{A}\tilde{\mathbf{u}} - \mathbf{b}\end{array}\right.$$

Rigal-Gaches (1967) proof

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \dots & \mathbf{A}_{1k} \\ \vdots & \dots & \vdots \\ \mathbf{A}_{p1} & \dots & \mathbf{A}_{pk} \end{bmatrix} \qquad \mathbf{r} = \begin{bmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_k \end{bmatrix} \qquad \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_p \end{bmatrix}$$

$$\Delta \mathbf{A}_{ij} = -\frac{(\mathbf{S}\|\tilde{\mathbf{u}}\|_{\vec{p}})_i}{(\mathbf{S}\|\tilde{\mathbf{u}}\|_{\vec{p}} + \|\mathbf{t}\|_{\vec{k}})_i} \mathbf{r}_i (\mathbf{Z}_j^i)^T$$

where

$$(\mathbf{Z}_j^i) = (\mathbf{S}\|\tilde{\mathbf{u}}\|_{\vec{p}})_i \mathbf{z}_k$$

and $\mathbf{z}_k$ is the dual vector of $\mathbf{u}_k$ ($\mathbf{z}_k^T \tilde{\mathbf{u}}_k = (\|\tilde{\mathbf{u}}\|_{\vec{p}})_k$;

$$\Delta \mathbf{b}_i = \frac{(\|\mathbf{t}\|_{\vec{k}})_i}{(\mathbf{S}\|\tilde{\mathbf{u}}\|_{\vec{p}} + \|\mathbf{t}\|_{\vec{k}})_i} \mathbf{r}_i$$

Backward error

We have the following equivalence in a general Hilbert (true also for a Banach):

$$\left.\begin{array}{l} \exists b \in \mathcal{BL}(\mathfrak{H}), \exists \delta L \in \mathfrak{H}^\star \text{ such that:}\\ a(\tilde{u}, v) + b(\tilde{u}, \mathbf{v}) = (L + \delta L)(v),\\ \forall v \in \mathfrak{H}, \ \text{and}\\ \|b(\cdot,\cdot)\|_{\mathcal{BL}(\mathfrak{H})} \le \alpha, \|\delta L\|_{\mathfrak{H}^\star} \le \beta \end{array}\right\} \Leftrightarrow \begin{cases} \|\rho_{\tilde{u}}\|_{\mathfrak{H}^\star} \le \alpha \|\tilde{u}\|_{\mathfrak{H}} + \beta \\ \text{where } \rho_{\tilde{u}} \in \mathfrak{H}^\star \text{ is defined by} \\ \langle \rho_{\tilde{u}}, v \rangle_{\mathfrak{H}^\star, \mathfrak{H}} = a(\tilde{u}, v) - L(v), \\ \forall v \in \mathfrak{H} \end{cases}$$

A., Noulard, and Russo (2001)

Backward error (proof)

The proof will be given assuming that $\mathfrak{H}$ is only a Banach space, thereby showing that the theorem holds, even in a more general situation. For this reason, in this proof, (and only here), we will use the notation of duality pairs.
$\Rightarrow$: This is obvious.
$\Leftarrow$: We will build two perturbations of a and L, respectively b and δL, such that :

$$a(\tilde{u}, v) + b(\tilde{u}, v) = L(v) + \delta L(v), \forall v \in \mathfrak{H}.$$

We set:

$$\forall u \in \mathfrak{H}, \langle \rho_u, v \rangle_{\mathfrak{H}^\star, \mathfrak{H}} = b(u, v) - L(v), \forall v \in \mathfrak{H};$$

we have $\rho_u \in \mathfrak{H}^\star$.

Backward error (proof)

We will denote by $J_u \in (\mathfrak{H}^\star)^\star = \mathfrak{H}^{\star\star}$ the element of the bi-dual of $\mathfrak{H}$, which is associated to u in the canonic injection

$$\begin{array}{rcccl} J & : & \mathfrak{H} & \longrightarrow & V_I^{\star\star} \subset \mathfrak{H}^{\star\star} \\ & & u & \longmapsto & J_u \end{array}$$

defined by $\langle J_u, f\rangle_{\mathfrak{H}^{\star\star},\mathfrak{H}^\star} = \langle f, u\rangle_{\mathfrak{H}^\star,\mathfrak{H}}, \forall f \in \mathfrak{H}^\star$. It is well-known that J is a linear isometry (see e.g. H. BRÉZIS, *Analyse Fonctionnelle, Théorie et Applications*, Masson, Paris, 1983.[III.4 p. 39]). We then have

$$\|J_{\tilde{u}}\|_{\mathfrak{H}^{\star\star}} = \|\tilde{u}\|_{\mathfrak{H}} = \sup_{\|f\|_{\mathfrak{H}^\star}\leq 1} \langle J_{\tilde{u}}, f\rangle_{\mathfrak{H}^{\star\star},\mathfrak{H}^\star} = \\ \sup_{\|f\|_{\mathfrak{H}^\star}\leq 1} \langle f, \tilde{u}\rangle_{\mathfrak{H}^\star,\mathfrak{H}} = \langle f_{\tilde{u}}, \tilde{u}\rangle_{\mathfrak{H}^\star,\mathfrak{H}},$$

for a certain $f_{\tilde{u}} \in \mathfrak{H}^\star$. One must be aware of the fact that, here, we cannot associate a vector $v \in \mathfrak{H}$ to $f_{\tilde{u}}$, unless $\mathfrak{H}$ is reflexive. In other words we cannot find a $v \in \mathfrak{H}$ such that $\|f_{\tilde{u}}\|_{\mathfrak{H}^\star} = \langle f_{\tilde{u}}, v\rangle_{\mathfrak{H}^\star,\mathfrak{H}}$, because $\|f_{\tilde{u}}\|_{\mathfrak{H}^\star}$ is a sup and not a max. It is a max if (and only if) $\mathfrak{H}$ is reflexive.

Backward error (proof)

Now, as has been done for the perturbation of a system of linear equations, we define:

$$b(u,v) = -\frac{\alpha}{\alpha\|\tilde{u}\|_{\mathfrak{H}} + \beta} \langle J_u, f_{\tilde{u}} \rangle_{\mathfrak{H}^{\star\star},\mathfrak{H}} \langle \rho_{\tilde{u}}, v \rangle_{\mathfrak{H}^{\star},\mathfrak{H}}$$

and

$$\delta L(v) = \frac{\beta}{\alpha\|\tilde{u}\|_{\mathfrak{H}} + \beta} \langle \rho_{\tilde{u}}, v \rangle_{\mathfrak{H}^{\star},\mathfrak{H}} .$$

It is obvious that b is continuous and bilinear from $\mathfrak{H} \times \mathfrak{H}$ to R, and $\delta L \in \mathfrak{H}^{\star}$; an easy computation shows that

$$\delta L(v) - b(\tilde{u}, v) =$$
$$\left(\frac{\beta}{\alpha\|\tilde{u}\|_{\mathfrak{H}} + \beta} + \frac{\alpha}{\alpha\|\tilde{u}\|_{\mathfrak{H}} + \beta} \langle J_{\tilde{u}}, f_{\tilde{u}} \rangle_{\mathfrak{H}^{\star\star},\mathfrak{H}^{\star}} \right) \langle \rho_{\tilde{u}}, v \rangle_{\mathfrak{H}^{\star},\mathfrak{H}} = \langle \rho_{\tilde{u}}, v \rangle_{\mathfrak{H}^{\star},\mathfrak{H}}$$

as required. Moreover, if we suppose that $\|\rho_{\tilde{u}}\|_{\mathfrak{H}^{\star}} \leq \alpha\|\tilde{u}\|_{\mathfrak{H}} + \beta$, then we have:

$$\|b\|_{\mathcal{BL}(\mathfrak{H})} \leq \alpha, \quad \|\delta L\|_{\mathfrak{H}^{\star}} \leq \beta.$$

Backward error (Remark)

If $\mathfrak{H}$ is a reflexive Banach space, we can give a more expressive form to the perturbation term b. In fact, in this case, we can identify J_u and u and obtain that

$$\begin{aligned} b(u,v) &= -\tfrac{\alpha}{\alpha\|\tilde{u}\|+\beta}\left\langle J_u, f_{\tilde{u}}\right\rangle_{\mathfrak{H}^{\star\star},\mathfrak{H}^{\star}}\left\langle \rho_{\tilde{u}}, v\right\rangle_{\mathfrak{H}^{\star},\mathfrak{H}} \\ &= -\tfrac{\alpha}{\alpha\|\tilde{u}\|+\beta}\left\langle f_{\tilde{u}}, u\right\rangle_{\mathfrak{H}^{\star},\mathfrak{H}}\left\langle \rho_{\tilde{u}}, v\right\rangle_{\mathfrak{H}^{\star},\mathfrak{H}} \\ &= -\tfrac{\alpha}{\alpha\|\tilde{u}\|+\beta}\left\langle f_{\tilde{u}}\otimes\rho_{\tilde{u}}, (u,v)\right\rangle, \end{aligned}$$

in analogy with the finite dimensional case.

The symmetric case: conjugate gradient method

$\mathbf{A}$ symmetric positive definite

$\mathfrak{H} = (\mathrm{R}^N, || \cdot ||_{\mathbf{A}})$ and $\mathfrak{H}^\star = (\mathrm{R}^N, || \cdot ||_{\mathbf{A}^{-1}})$

At each step k the conjugate gradient method minimizes the energy norm of the error $\delta \mathbf{u}^{(k)} = \mathbf{u} - \mathbf{u}^{(k)}$ on a Krylov space $\mathbf{u}^{(0)} + \mathcal{K}_k$:

$$\min_{\mathbf{u}^{(k)} \in \, \mathbf{u}^{(0)} + \mathcal{K}_k} \|\delta \mathbf{u}^{(k)}\|_{\mathbf{A}}^2$$

$$\|\delta \mathbf{u}^{(k)}\|_{\mathbf{A}} = \|\rho_{\mathbf{u}^{(k)}}\|_{\mathfrak{H}^\star} = \|\widehat{\mathbf{r}}^{(k)}\|_{\mathbf{A}^{-1}}$$

$$\widehat{\mathbf{r}}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{u}^{(k)}$$

Arioli Numer. Math. (2003)

The symmetric case: conjugate gradient method

$$
\begin{aligned}
\mathbf{u}^{(k)} &= \mathbf{u}^{(k-1)} + \alpha_{k-1}\mathbf{p}^{(k-1)} & \alpha_{k-1} &= \frac{\mathbf{r}^{(k-1)T}\mathbf{r}^{(k-1)}}{\mathbf{p}^{(k-1)T}\mathbf{A}\mathbf{p}^{(k-1)}}, \\
\mathbf{r}^{(k)} &= \mathbf{r}^{(k-1)} - \alpha_{k-1}\mathbf{A}\mathbf{p}^{(k-1)} & & \\
\mathbf{p}^{(k)} &= \mathbf{r}^{(k)} + \beta_{k-1}\mathbf{p}^{(k-1)}, & \beta_{k-1} &= \frac{\mathbf{r}^{(k)T}\mathbf{r}^{(k)}}{\mathbf{r}^{(k-1)T}\mathbf{r}^{(k-1)}},
\end{aligned}
$$

where $\mathbf{u}^{(0)} = 0$ and $\mathbf{r}^{(0)} = \mathbf{p}^{(0)} = \mathbf{b}$.

The symmetric case: conjugate gradient method

Taking into account that $\mathbf{p}^{(i)T}\mathbf{A}\mathbf{p}^{(j)} = 0, i \neq j$ we have

$$\mathbf{u} = \sum_{j=1}^{N} \alpha_j \mathbf{p}^{(j)} \qquad \|\mathbf{u}\|_{\mathbf{A}}^2 = \sum_{j=1}^{N} \alpha_j \mathbf{r}^{(j)T}\mathbf{r}^{(j)}$$

The symmetric case: conjugate gradient method

Taking into account that $\mathbf{p}^{(i)\,T}\mathbf{A}\mathbf{p}^{(j)} = 0, i \neq j$ we have

$$\mathbf{u} = \sum_{j=1}^{N} \alpha_j \mathbf{p}^{(j)} \qquad \|\mathbf{u}\|_{\mathbf{A}}^2 = \sum_{j=1}^{N} \alpha_j \mathbf{r}^{(j)\,T}\mathbf{r}^{(j)}$$

$$\begin{aligned} \mathbf{u}^T\mathbf{A}\mathbf{u} &= \textstyle\sum_{j=1}^{N}\sum_{i=1}^{N} \alpha_j\alpha_i \mathbf{p}^{(j)\,T}\mathbf{A}\mathbf{p}^{(i)} \\ &= \textstyle\sum_{j=1}^{N} \alpha_j^2 \mathbf{p}^{(j)\,T}\mathbf{A}\mathbf{p}^{(j)} \end{aligned}$$

$$\text{but} \quad \alpha_j \mathbf{p}^{(j)\,T}\mathbf{A}\mathbf{p}^{(j)} = \mathbf{r}^{(j)\,T}\mathbf{r}^{(j)}.$$

The symmetric case: stopping criteria

- Classic Criterion:

$$\text{IF} \quad \|\mathbf{Au}^{(k)} - \mathbf{b}\|_2 \leq \sqrt{\varepsilon}\|\mathbf{b}\|_2 \quad \text{THEN STOP} \; ,$$

The symmetric case: stopping criteria

- Classic Criterion:

$$\text{IF} \;\; \|\mathbf{A}\mathbf{u}^{(k)} - \mathbf{b}\|_2 \leq \sqrt{\varepsilon}\|\mathbf{b}\|_2 \;\; \text{THEN STOP} \; ,$$

- New Criterion:

$$\text{IF} \;\; \|\mathbf{A}\mathbf{u}^{(k)} - \mathbf{b}\|_{\mathbf{A}^{-1}} \leq \eta\|\mathbf{b}\|_{A^{-1}} \;\; \text{THEN STOP} \; ,$$

with $\eta < 1$ an a-priori threshold fixed by the user.

The symmetric case: stopping criteria

- Classic Criterion:

$$\text{IF} \quad \|\mathbf{Au}^{(k)} - \mathbf{b}\|_2 \leq \sqrt{\varepsilon}\|\mathbf{b}\|_2 \quad \text{THEN STOP} \ ,$$

- New Criterion:

$$\text{IF} \quad \|\mathbf{Au}^{(k)} - \mathbf{b}\|_{\mathbf{A}^{-1}} \leq \eta\|\mathbf{b}\|_{A^{-1}} \quad \text{THEN STOP} \ ,$$

with $\eta < 1$ an a-priori threshold fixed by the user. The choice of η will depend on the properties of the problem that we want to solve, and, in the practical cases, η can be frequently much larger than ε , the roundoff unit of the computer finite precision arithmetic.

The symmetric case: stopping criteria cont.

$$\|\mathbf{A}\mathbf{u}^{(k)} - \mathbf{b}\|_{\mathbf{A}^{-1}} \quad ?$$

The symmetric case: stopping criteria cont.

$$\|\mathbf{A}\mathbf{u}^{(k)} - \mathbf{b}\|_{\mathbf{A}^{-1}} \quad ?$$

$$\|\mathbf{b}\|_{A^{-1}} \quad ?$$

The symmetric case: stopping criteria cont.

$$\|\mathbf{A}\mathbf{u}^{(k)} - \mathbf{b}\|_{\mathbf{A}^{-1}}$$

The symmetric case: stopping criteria cont.

$$\|\mathbf{A}\mathbf{u}^{(k)} - \mathbf{b}\|_{\mathbf{A}^{-1}}$$

- Hestenes-Stiefel rule (1952) (see Strakoš and Tichý, 2002) numerically stable

The symmetric case: stopping criteria cont.

$$\|\mathbf{A}\mathbf{u}^{(k)} - \mathbf{b}\|_{\mathbf{A}^{-1}}$$

- Hestenes-Stiefel rule (1952) (see Strakoš and Tichý, 2002) numerically stable
- Gauss quadrature rules (Golub and Meurant, 1997)

The symmetric case: stopping criteria cont.

$$\|\mathbf{Au}^{(k)} - \mathbf{b}\|_{\mathbf{A}^{-1}}$$

- Hestenes-Stiefel rule (1952) (see Strakoš and Tichý, 2002) numerically stable
- Gauss quadrature rules (Golub and Meurant, 1997)
  - Gauss equivalent to Hestenes-Stiefel rule (Strakoš and Tichý). The Gauss quadrature does not require any a-priori knowledge of the smallest and the biggest eigenvalues and computes a lower bound of$\|\mathbf{Au}^{(k)} - \mathbf{b}\|_{\mathbf{A}^{-1}}$.

The symmetric case: stopping criteria cont.

$$\|\mathbf{Au}^{(k)} - \mathbf{b}\|_{\mathbf{A}^{-1}}$$

- Hestenes-Stiefel rule (1952) (see Strakoš and Tichý, 2002) numerically stable
- Gauss quadrature rules (Golub and Meurant, 1997)
  - Gauss equivalent to Hestenes-Stiefel rule (Strakoš and Tichý). The Gauss quadrature does not require any a-priori knowledge of the smallest and the biggest eigenvalues and computes a lower bound of$\|\mathbf{Au}^{(k)} - \mathbf{b}\|_{\mathbf{A}^{-1}}$.
  - Gauss-Lobatto and Gauss-Radau. They compute lower and upper bounds using the extremes eigenvalues of $\mathbf{A}$.

The symmetric case: Hestenes-Stiefel rule

During the conjugate gradient iterates, we compute the scalar α_k and the conjugate vectors $\mathbf{p}^{(k)}$ ($\mathbf{p}^{(j)\,T}\mathbf{A}\mathbf{p}^{(i)} = 0,\ j \neq i$) and the residuals $\mathbf{r}^{(k)}$. Thus,

$$\mathbf{u} = \sum_{j=1}^{\mathbf{N}} \alpha_j \mathbf{p}^{(j)}$$

and

$$\|\delta \mathbf{u}^{(k)}\|_{\mathbf{A}}^2 = \|\mathbf{A}\mathbf{u}^{(k)} - \mathbf{b}\|_{\mathbf{A}^{-1}}^2 = e_{\mathbf{A}}^2 = \sum_{j=k+1}^{\mathbf{N}} \alpha_j \mathbf{r}^{(j)\,T}\mathbf{r}^{(j)}$$

The symmetric case: Hestenes-Stiefel rule

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 ξ_k will be

$$\xi_k = \sum_{j=k+1}^{k+d} \alpha_j \mathbf{r}^{(j)T} \mathbf{r}^{(j)}.$$

The symmetric case: Hestenes-Stiefel rule

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 ξ_k will be

$$\xi_k = \sum_{j=k+1}^{k+d} \alpha_j \mathbf{r}^{(j)T} \mathbf{r}^{(j)}.$$

The choice of a value for d depends on preconditioner and ill-conditioning.

$\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}$

From

$$\mathbf{r}^{(k)\,T}\mathbf{v} = 0, \qquad \forall \mathbf{v} \in \mathcal{K}_k,$$

we prove

$$\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} = \mathbf{u}^T\mathbf{A}\mathbf{u} \geq \sum_{j=1}^{k} \alpha_j \mathbf{r}^{(j)\,T}\mathbf{r}^{(j)},$$

(the right-hand side will converge monotonically to $\|\mathbf{u}\|^2_{\mathbf{A}}$).
Therefore, we use the following stopping criterion

$$\text{IF} \;\; \xi_k \leq \eta^2 \sum_{j=1}^{k} \alpha_j \mathbf{r}^{(j)\,T}\mathbf{r}^{(j)} \;\; \text{THEN STOP} .$$

Preconditioning

Let $\mathbf{U}$ a non singular matrix: the symmetric preconditioned system is

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

$$\begin{aligned}
\mathbf{y}^{(k)} &= \mathbf{y}^{(k-1)} + \alpha_{k-1}\widehat{\mathbf{p}}^{(k-1)} & \alpha_{k-1} &= \frac{\widehat{\mathbf{r}}^{(k-1)\,T}\widehat{\mathbf{r}}^{(k-1)}}{\widehat{\mathbf{p}}^{(k-1)\,T}\mathbf{U}^{-T}\mathbf{A}\mathbf{U}^{-1}\widehat{\mathbf{p}}^{(k-1)}},\\
\widehat{\mathbf{r}}^{(k)} &= \widehat{\mathbf{r}}^{(k-1)} - \alpha_{k-1}\mathbf{U}^{-T}\mathbf{A}\mathbf{U}^{-1}\widehat{\mathbf{p}}^{(k-1)} & & \\
\widehat{\mathbf{p}}^{(k)} &= \widehat{\mathbf{r}}^{(k)} + \beta_{k-1}\widehat{\mathbf{p}}^{(k-1)}, & \beta_{k-1} &= \frac{\widehat{\mathbf{r}}^{(k)\,T}\widehat{\mathbf{r}}^{(k)}}{\widehat{\mathbf{r}}^{(k-1)\,T}\widehat{\mathbf{r}}^{(k-1)}},
\end{aligned}$$

where $\mathbf{y}^{(0)} = 0$ and $\widehat{\mathbf{r}}^{(0)} = \widehat{\mathbf{p}}^{(0)} = \mathbf{b}$. In exact arithmetic we have

$$\widehat{\mathbf{r}}^{(k)} = \mathbf{U}^{-T}\mathbf{b} - \mathbf{U}^{-T}\mathbf{A}\mathbf{U}^{-1}\mathbf{y}^{(k)}.$$

Preconditioning

Let $\mathbf{U}$ a non singular matrix: the symmetric preconditioned system is

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

$$\begin{aligned}
\mathbf{y}^{(k)} &= \mathbf{y}^{(k-1)} + \alpha_{k-1}\widehat{\mathbf{p}}^{(k-1)} & \alpha_{k-1} &= \frac{\widehat{\mathbf{r}}^{(k-1)T}\widehat{\mathbf{r}}^{(k-1)}}{\widehat{\mathbf{p}}^{(k-1)T}\mathbf{U}^{-T}\mathbf{A}\mathbf{U}^{-1}\widehat{\mathbf{p}}^{(k-1)}}, \\
\widehat{\mathbf{r}}^{(k)} &= \widehat{\mathbf{r}}^{(k-1)} - \alpha_{k-1}\mathbf{U}^{-T}\mathbf{A}\mathbf{U}^{-1}\widehat{\mathbf{p}}^{(k-1)} & & \\
\widehat{\mathbf{p}}^{(k)} &= \widehat{\mathbf{r}}^{(k)} + \beta_{k-1}\widehat{\mathbf{p}}^{(k-1)}, & \beta_{k-1} &= \frac{\widehat{\mathbf{r}}^{(k)T}\widehat{\mathbf{r}}^{(k)}}{\widehat{\mathbf{r}}^{(k-1)T}\widehat{\mathbf{r}}^{(k-1)}},
\end{aligned}$$

where $\mathbf{y}^{(0)} = 0$ and $\widehat{\mathbf{r}}^{(0)} = \widehat{\mathbf{p}}^{(0)} = \mathbf{b}$. In exact arithmetic we have

$$\widehat{\mathbf{r}}^{(k)} = \mathbf{U}^{-T}\mathbf{b} - \mathbf{U}^{-T}\mathbf{A}\mathbf{U}^{-1}\mathbf{y}^{(k)}.$$

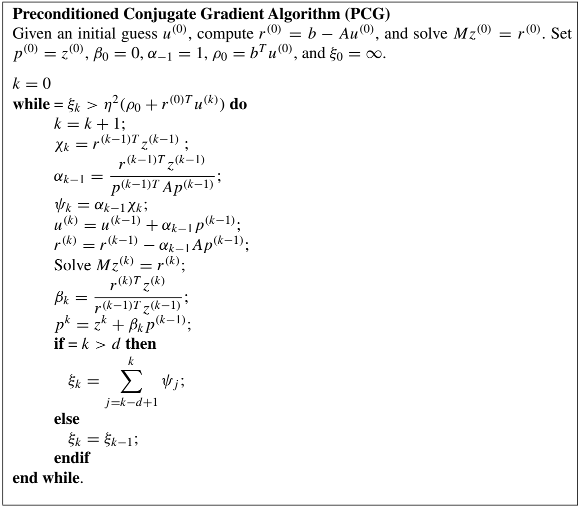
Defining $\mathbf{u}^{(k)} = \mathbf{U}^{-1}\mathbf{y}^{(k)}$ we have $\widehat{\mathbf{r}}^{(k)} = \mathbf{U}^{-T}\mathbf{r}^{(k)}$. Then

$$\|\widehat{\mathbf{r}}^{(k)}\|^2_{(\mathbf{U}^{-T}\mathbf{A}\mathbf{U}^{-1})^{-1}} = \widehat{\mathbf{r}}^{(k)T}\mathbf{U}\mathbf{A}^{-1}\mathbf{U}^{T}\widehat{\mathbf{r}}^{(k)} = \|\mathbf{r}^{(k)}\|^2_{\mathbf{A}^{-1}}$$

Preconditioning

The dual norm of the preconditioned residual is equal to the dual norm of the original residual.

PCG algorithm



Preconditioned Conjugate Gradient Algorithm (PCG)

Given an initial guess $u^{(0)}$, compute $r^{(0)} = b - Au^{(0)}$, and solve $Mz^{(0)} = r^{(0)}$. Set $p^{(0)} = z^{(0)}$, $\beta_0 = 0$, $\alpha_{-1} = 1$, $\rho_0 = b^T u^{(0)}$, and $\xi_0 = \infty$.

```
$k = 0$
**while** = $\xi_k > \eta^2(\rho_0 + r^{(0)T} u^{(k)})$ **do**
    $k = k + 1$;
    $\chi_k = r^{(k-1)T} z^{(k-1)}$ ;
    $\alpha_{k-1} = \frac{r^{(k-1)T} z^{(k-1)}}{p^{(k-1)T} A p^{(k-1)}}$;
    $\psi_k = \alpha_{k-1} \chi_k$;
    $u^{(k)} = u^{(k-1)} + \alpha_{k-1} p^{(k-1)}$;
    $r^{(k)} = r^{(k-1)} - \alpha_{k-1} A p^{(k-1)}$;
    Solve $Mz^{(k)} = r^{(k)}$;
    $\beta_k = \frac{r^{(k)T} z^{(k)}}{r^{(k-1)T} z^{(k-1)}}$;
    $p^k = z^k + \beta_k p^{(k-1)}$;
    **if** = $k > d$ **then**
        $\xi_k = \sum_{j=k-d+1}^{k} \psi_j$;
    **else**
        $\xi_k = \xi_{k-1}$;
    **endif**
**end while**.
```

Fig. 1. Preconditioned Conjugate Gradient Algorithm (PCG)

Continuous problem

$$a(u, v) = \int_\Omega \mathfrak{k}(\mathbf{x}) \nabla u \cdot \nabla v d\mathbf{x}, \quad \forall u, v \in H_0^1(\Omega)$$

$\forall u, v \in H_0^1(\Omega)$, $\exists \gamma \in \mathbb{R}_+$ and $\exists M \in \mathbb{R}_+$ such that

$$\begin{aligned} \gamma ||u||_{1,\Omega}^2 \leq \quad & a(u, u) \\ & a(u, v) \quad \leq M ||u||_{1,\Omega} ||v||_{1,\Omega} \ , \end{aligned}$$

$L(v) = \int_\Omega f v d\mathbf{x}$, $L(v) \in H^{-1}(\Omega)$.

$$(P) \left\{ \begin{array}{l} \text{Find } u \in H_0^1(\Omega) \text{ such that} \\ a(u, v) = L(v), \ \ \forall v \in H_0^1(\Omega), \end{array} \right. \qquad \text{has a unique solution.}$$

Finite-element approximation

- Weak formulation

Finite-element approximation

- Weak formulation

$$\begin{cases} \text{Find } u_h \in \mathfrak{H}_h \text{ such that} \\ a_h(u_h, v_h) = L_h(v_h), \\ \forall v_h \in \mathfrak{H}_h, \end{cases}$$

Finite element methods choose $\mathfrak{H}_h$ to be a space of functions v_h defined on a subdivision Ω_h of Ω into simplices T of diameter h_T ; h denotes a piecewise constant function defined on Ω_h via $h|_T = h_T$.

Finite-element approximation

- Weak formulation

$$\left\{\begin{array}{l}\text{Find } u_h \in \mathfrak{H}_h \text{ such that}\\ a_h(u_h, v_h) = L_h(v_h),\\ \forall v_h \in \mathfrak{H}_h,\end{array}\right.$$

  Finite element methods choose $\mathfrak{H}_h$ to be a space of functions v_h defined on a subdivision Ω_h of Ω into simplices T of diameter h_T ; h denotes a piecewise constant function defined on Ω_h via $h|_T = h_T$.
- Existence and uniqueness: $\mathfrak{H}_h \subset \mathfrak{H} = H_0^1(\Omega)$.
- Error Estimate: $\|u - u_h\|_{\mathfrak{H}} \leq C(h)$

  See Claes Johnson Numerical Solutions Of Partial Differential Equations By The Finite Element Method

  2009

Finite-element framework

Solve

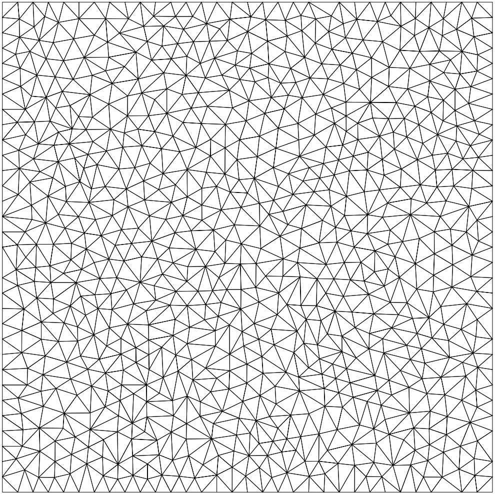
$$\mathbf{A}\mathbf{u}_h = \mathbf{b}$$

given

$$\begin{aligned}
\sup_{\mathbf{w}\in\mathbb{R}^N\setminus\{\mathbf{0}\}} \sup_{\mathbf{v}\in\mathbb{R}^N\setminus\{\mathbf{0}\}} \frac{\mathbf{w}^T\mathbf{A}\mathbf{v}}{\|\mathbf{v}\|_{\mathbf{H}}\|\mathbf{w}\|_{\mathbf{H}}} &\leq C_1 && \text{(sup-sup)}\\
\inf_{\mathbf{w}\in\mathbb{R}^N\setminus\{\mathbf{0}\}} \sup_{\mathbf{v}\in\mathbb{R}^N\setminus\{\mathbf{0}\}} \frac{\mathbf{w}^T\mathbf{A}\mathbf{v}}{\|\mathbf{v}\|_{\mathbf{H}}\|\mathbf{w}\|_{\mathbf{H}}} &\geq C_2 && \text{(inf-sup)}
\end{aligned}$$

Note: $\|v_h\|_{\mathfrak{H}_h} = \|\mathbf{v}\|_{\mathbf{H}}$.

Example: Mesh



Finite-element framework

Finally, assuming $h < 1$ and $t > 0$, and choosing $\eta = \mathcal{O}(h)$, we have

$$\|u - u_h^{(k)}\|_{\mathfrak{H}} \leq C^*(h^t)\|u\|_{\mathfrak{H}} + 2\|u - u_h\|_{\mathfrak{H}} \leq C(h).$$

where

- $u(\mathbf{x})$ is the exact solution of the variational problem,
- $u_h(\mathbf{x})$ is the exact solution of the approximate problem,
- $u_h^{(k)}(\mathbf{x}) = \sum_{i=1}^N \mathbf{u}_h^{(k)}\phi_i(\mathbf{x})$ is the approximate solution at step k. ($\phi_i(\mathbf{x})$ are the basis functions)

Test problems

Problem 1

$$\mathfrak{k}(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in \Omega \setminus \{\Omega_1 \cup \Omega_2 \cup \Omega_3\}, \\ 10^{-6} & \mathbf{x} \in \Omega_1, \\ 10^{-4} & \mathbf{x} \in \Omega_2, \\ 10^{-2} & \mathbf{x} \in \Omega_3. \end{cases}$$

Problem 2

$$\mathfrak{k}(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in \Omega \setminus \{\Omega_1 \cup \Omega_2 \cup \Omega_3\}, \\ 10^{6} & \mathbf{x} \in \Omega_1, \\ 10^{4} & \mathbf{x} \in \Omega_2, \\ 10^{2} & \mathbf{x} \in \Omega_3. \end{cases}$$

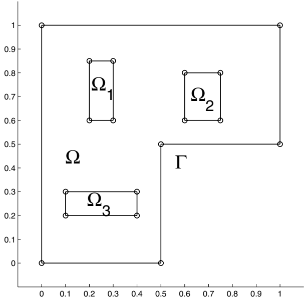


Fig. 2. Geometry of the domain Ω

$$L(v) = \int_\Omega 10 v d\mathbf{x}, \qquad \forall v \in H^1_0(\Omega)$$

Preconditioners: estimates for $\kappa(\mathbf{M}^{-1}\mathbf{A})$

M	Problem 1	Problem 2
I	$3.6\ 10^{8}$	$1.8\ 10^{10}$
Jacobi	$2.4\ 10^{4}$	$1.5\ 10^{9}$
Inc. Cholesky(0)	$7.2\ 10^{3}$	$4.3\ 10^{8}$

$$\eta^2 = 3.44.30510^{-5} \text{ and } \mathbf{N} = 29619.$$

The condition numbers of the preconditioned matrices $\mathbf{M}^{-1}\mathbf{A}$ for the second problem are are still very high, and only the incomplete Cholesky preconditioner with drop tolerance 10^{-2} is an effective choice.

Example: Problem 1

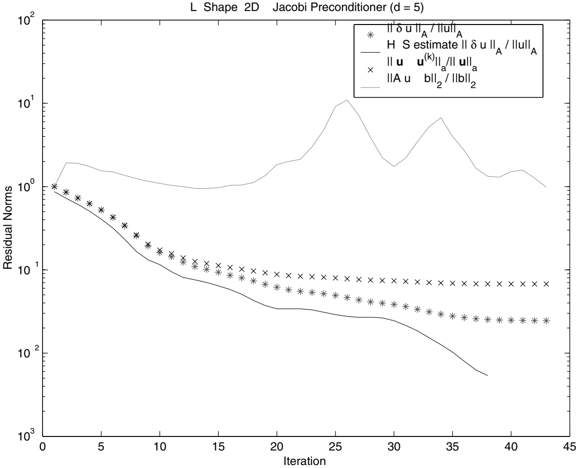


Fig. 3. Behaviour of the norms of the residual for the Jacobi preconditioner in Problem 1

Behaviour of the norms of the residual for the Jacobi

Example: Problem 1

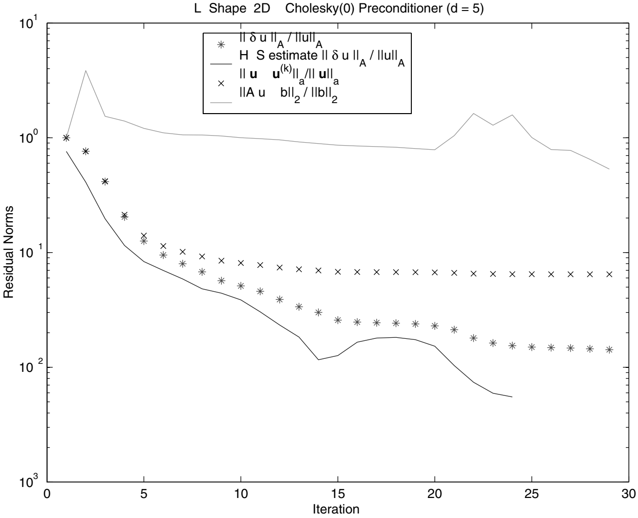


Fig. 4. Behaviour of the norms of the residual for the incomplete Cholesky preconditioner in Problem 1

Example: Problem 2

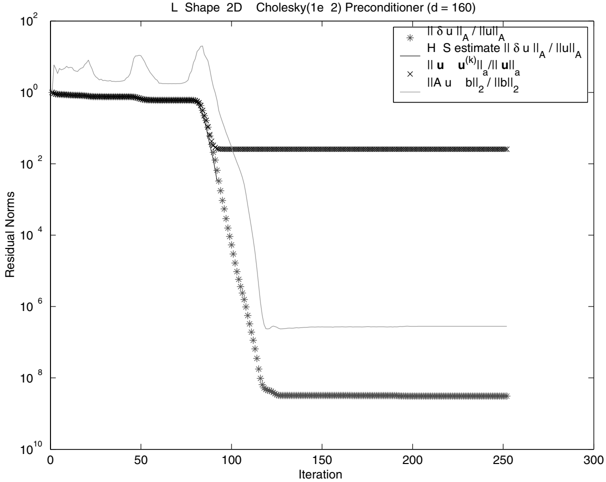


Fig. 10. Behaviour of the norms of the residual for the incomplete Cholesky preconditioner with drop tolerance 10^{-2} and $d = 160$ in Problem 2

Example: Problem 2

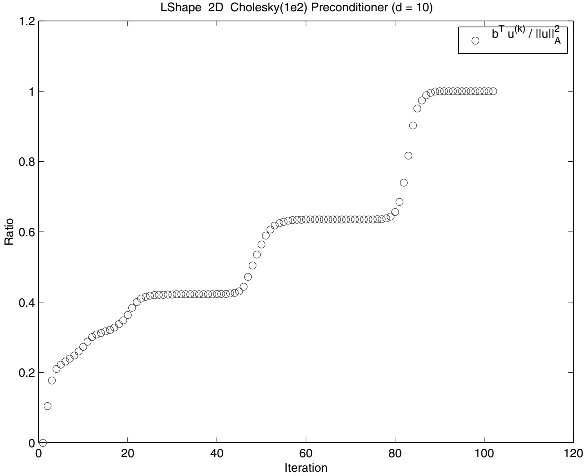


Fig. 9. Ratio $b^T u^{(k)}/||u||_A^2$ for the incomplete Cholesky preconditioner with drop tolerance 10^{-2} and $d = 10$ in Problem 2

Example: Problem 2

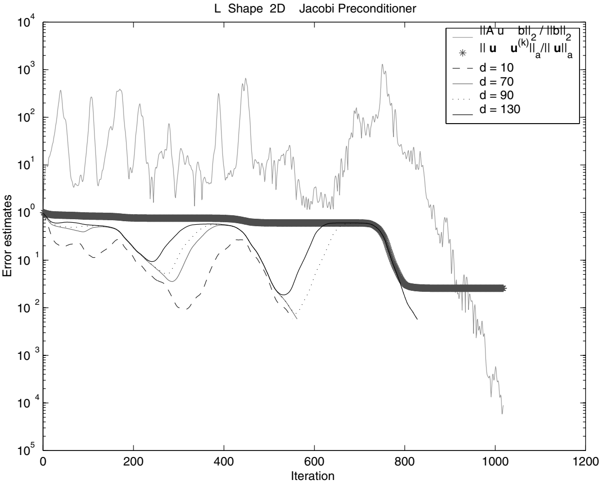


Fig. 6. Comparison of several estimates of the energy error for $d = 10, 70, 90, 130$ in Problem 2

Lecture 3

The non symmetric positive definite problem

- $a(u, v) \neq a(v, u)$
- $\mathbf{A}$ asymmetric but positive definite
- $\mathbf{H} = \frac{1}{2}(\mathbf{A}^T + \mathbf{A})$ SPD
- $\mathbf{A} = \frac{1}{2}(\mathbf{A}^T + \mathbf{A}) + \frac{1}{2}(\mathbf{A}^T - \mathbf{A}) = \mathbf{H} - \mathbf{N}$

How to calculate $\|\mathbf{r}^{(k)}\|_{\mathbf{H}^{-1}}$?

- Solve preconditioned system

$$\mathbf{H}^{-1/2}\mathbf{A}\mathbf{H}^{-1/2}\hat{\mathbf{u}} = \mathbf{H}^{-1/2}\mathbf{b}$$

  - $\|\hat{\mathbf{r}}^{(k)}\|_{l_2} = \|\mathbf{r}^{(k)}\|_{\mathbf{H}^{-1}}$
  - 3-term recurrence
- Approximate it from Krylov subspace information.

See A., Login, and Wathen

Numer. Math. (2004) (DOI) 10.1007/s00211-004-0568-z

A. and Loghin Electronic Transactions on Numerical Analysis. 29, (2008).

inf-sup framework

Solve

$$\mathbf{A}\mathbf{u} = \mathbf{f}$$

given

$$\begin{aligned}\sup_{\mathbf{w}\in\mathrm{R}^n\setminus\{\mathbf{0}\}}\ \sup_{\mathbf{v}\in\mathrm{R}^n\setminus\{\mathbf{0}\}} \frac{\mathbf{w}^T\mathbf{A}\mathbf{v}}{\|\mathbf{v}\|_{\mathbf{H}}\|\mathbf{w}\|_{\mathbf{H}}} &\leq C_1 & \text{(sup-sup)}\\ \inf_{\mathbf{w}\in\mathrm{R}^n\setminus\{\mathbf{0}\}}\ \sup_{\mathbf{v}\in\mathrm{R}^n\setminus\{\mathbf{0}\}} \frac{\mathbf{w}^T\mathbf{A}\mathbf{v}}{\|\mathbf{v}\|_{\mathbf{H}}\|\mathbf{w}\|_{\mathbf{H}}} &\geq C_2 & \text{(inf-sup)}\end{aligned}$$

Note: $\|v_h\|_{\mathfrak{H}_h} = \|\mathbf{v}\|_{\mathbf{H}}$ defines the spd matrix $\mathbf{H}$.

3-term recurrence Algorithm

$$\mathcal{K}_k = \mathrm{span}\left\{\mathbf{H}^{-1}\mathbf{f}, \mathbf{H}^{-1}\mathbf{N}\mathbf{f}, \ldots, \left(\mathbf{H}^{-1}\mathbf{N}\right)^{k-1}\mathbf{f}\right\}$$

3-term recurrence Algorithm

$$\mathcal{K}_k = \mathrm{span}\left\{\mathbf{H}^{-1}\mathbf{f}, \mathbf{H}^{-1}\mathbf{N}\mathbf{f}, \ldots, \left(\mathbf{H}^{-1}\mathbf{N}\right)^{k-1}\mathbf{f}\right\}$$

We compute the Lanczos vectors $\mathbf{v}^{(j)}$ by a 3-term recurrence:

$$\alpha_j \mathbf{v}^{(j+1)} = \mathbf{H}^{-1}\mathbf{N}\mathbf{v}^{(j)} - \gamma_j \mathbf{v}^{(j)} - \beta_j \mathbf{v}^{(j-1)}, \qquad j \geq 0$$

with $\mathbf{v}^{(-1)} = 0$ and $\mathbf{v}^{(0)} = \mathbf{H}^{-1}\mathbf{N}\mathbf{f}$ The coefficients α_j, γ_j, and β_j are chosen such that

$$\mathbf{v}^{(i)\,T}\mathbf{H}\mathbf{v}^{(j)} = \delta_{ij}$$

i.e. they are **H** orthogonal. Widlund SINUM,15, 1978

3-term recurrence Algorithm

$$\mathcal{K}_k = \text{span}\left\{\mathbf{H}^{-1}\mathbf{f}, \mathbf{H}^{-1}\mathbf{N}\mathbf{f}, \ldots, \left(\mathbf{H}^{-1}\mathbf{N}\right)^{k-1}\mathbf{f}\right\}$$

We compute the Lanczos vectors $\mathbf{v}^{(j)}$ by a 3-term recurrence:

$$\alpha_j \mathbf{v}^{(j+1)} = \mathbf{H}^{-1}\mathbf{N}\mathbf{v}^{(j)} - \gamma_j \mathbf{v}^{(j)} - \beta_j \mathbf{v}^{(j-1)}, \qquad j \geq 0$$

with $\mathbf{v}^{(-1)} = 0$ and $\mathbf{v}^{(0)} = \mathbf{H}^{-1}\mathbf{N}\mathbf{f}$ The coefficients α_j, γ_j, and β_j are chosen such that

$$\mathbf{v}^{(i)T}\mathbf{H}\mathbf{v}^{(j)} = \delta_{ij}$$

i.e. they are **H** orthogonal. Widlund SINUM,15, 1978 This is possible only in this case for the peculiar preconditioning and the Skew-Symmetry of **N**. In general, we cannot have 3-term recurrent formulae for non-symmetric matrices (see Faber-Manteuffel SINUM, 21, 1984)

One crime

Replace

$$\|u - u_h\|_{\mathfrak{H}_h} \leq C(h)$$

with

$$\|u - u_h^{(k)}\|_{\mathfrak{H}_h} \leq C(h)$$

One crime

Replace

$$\|u - u_h\|_{\mathfrak{H}_h} \leq C(h)$$

with

$$\|u - u_h^{(k)}\|_{\mathfrak{H}_h} \leq C(h)$$

Sufficient condition

$$\|u - u_h\|_{\mathfrak{H}_h} + \|u_h - u_h^{(k)}\|_{\mathfrak{H}_h} \sim O(C(h))$$

$$\Downarrow$$

$$\|u_h - u_h^{(k)}\|_{\mathfrak{H}_h} \sim O(C(h))$$

Stopping criteria

A general stopping criterion:

$$\|u_h - u_h^{(k)}\|_{\mathfrak{H}_h} = \|\mathbf{u} - \mathbf{u}^{(k)}\|_{\mathbf{H}} \leq C(h)$$

Stopping criteria

A general stopping criterion:

$$\|u_h - u_h^{(k)}\|_{\mathfrak{H}_h} = \|\mathbf{u} - \mathbf{u}^{(k)}\|_{\mathbf{H}} \leq C(h)$$

Residual equation

$$\mathbf{r}^{(k)} = \mathbf{A}(\mathbf{u} - \mathbf{u}^{(k)})$$

$$\Downarrow$$

$$\|\mathbf{u} - \mathbf{u}^{(k)}\|_{\mathbf{H}} = \|\mathbf{A}^{-1}\mathbf{r}^{(k)}\|_{\mathbf{H}} = \|\mathbf{r}^{(k)}\|_{\mathbf{A}^{-T}\mathbf{H}\mathbf{A}^{-1}} \leq C(h)$$

Stopping criteria

Lemma *Let (inf-sup) hold. Then*

$$\|\mathbf{r}^{(k)}\|_{\mathbf{A}^{-T}\mathbf{H}\mathbf{A}^{-1}} \leq C_2^{-1}\|\mathbf{r}^{(k)}\|_{\mathbf{H}^{-1}}.$$

Stopping criteria

Lemma *Let (inf-sup) hold. Then*

$$\|\mathbf{r}^{(k)}\|_{\mathbf{A}^{-T}\mathbf{H}\mathbf{A}^{-1}} \leq C_2^{-1}\|\mathbf{r}^{(k)}\|_{\mathbf{H}^{-1}}.$$

New stopping criterion

$$\|\mathbf{r}^{(k)}\|_{\mathbf{H}^{-1}} \leq C_2 C(h)\|\mathbf{u}^{(k)}\|_{\mathbf{H}}.$$

Examples

Elliptic problems in $\mathbb{R}^2$ (Ω unit square)

$$\begin{aligned} -\nabla \cdot (\mathbf{a}(\mathbf{x}) \nabla u) + \mathbf{b}(\mathbf{x}) \cdot \nabla u + c(\mathbf{x}) u &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \Gamma. \end{aligned}$$

where

$$(\mathbf{a})_{ij},\ (\mathbf{b})_i,\ c \in L^\infty(\Omega), \qquad i, j = 1, 2,$$

$$k_2(\mathbf{x}) \left|\boldsymbol{\xi}\right|^2 \leq \boldsymbol{\xi}^T \mathbf{a}(\mathbf{x}) \boldsymbol{\xi} \leq k_1(\mathbf{x}) \left|\boldsymbol{\xi}\right|^2,$$

$$c(\mathbf{x}) - \frac{1}{2} \nabla \cdot \mathbf{b}(\mathbf{x}) \geq 0 \qquad \forall \mathbf{x} \in \Omega.$$

Examples

$$a(w, v) = (\mathbf{a} \cdot \nabla w, \nabla v) + (\mathbf{b} \cdot \nabla w, v) + (cw, v),$$

is continuous and coercive with

$$C_1 = \|k_1\|_{L^\infty(\Omega)} + \|\mathbf{b}\|_{L^\infty(\Omega)} + C(\Omega)\|c\|_{L^\infty(\Omega)},$$

$$C_2 = \min_{\mathbf{x}\in\Omega} k_2(\mathbf{x}),$$

wrt $\|\cdot\|_{\mathfrak{H}} = |\cdot|_{H_0^1(\Omega)} := |\cdot|_1.$

Examples

Error estimate:

$$|u - u_h|_1 \leq Ch^{s-1}\|u\|_s, \qquad 1 \leq s \leq 2.$$

Examples

Error estimate:

$$|u - u_h|_1 \leq Ch^{s-1}\|u\|_s, \qquad 1 \leq s \leq 2.$$

Issues

- What is h?
- How to approximate $\|u\|_s$?

Numerical experiments

- Discretization:
  linear elements on uniform & adaptive meshes

Numerical experiments

- Discretization:
  linear elements on uniform & adaptive meshes
- Estimation of parameters

$$h \sim \frac{\|\mathbf{u}^k\|_{\mathbf{M}}}{\|\mathbf{u}^k\|_{l_2}}, \quad \|u\|_s \sim \|\mathbf{A}\mathbf{u}^k\|_{l_2}$$

Numerical experiments

Stopping criteria and estimates

- Residual dual norm: $\|\mathbf{r}^k\|_{\mathbf{H}^{-1}}$

Numerical experiments

Stopping criteria and estimates

- Residual dual norm: $\|\mathbf{r}^k\|_{\mathbf{H}^{-1}}$
- Energy estimate $\|\mathbf{u}^k - \mathbf{u}^{k-1}\|_{\mathbf{H}} \leq C_2 h^2 \|\mathbf{A}\mathbf{u}^k\|_{l_2}$

Advection-diffusion problem

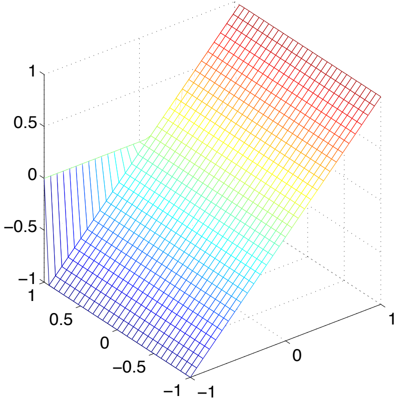
$$\begin{aligned} -\varepsilon \Delta u + \mathbf{b} \cdot \nabla u &= f \qquad \text{in } \Omega \\ u &= g \qquad \text{on } \Gamma. \end{aligned}$$

$$\mathbf{b} = (2y(1-x^2), -2x(1-y^2)),$$

$$u(x,y) = x \left(\frac{1 - e^{\frac{y-1}{\varepsilon}}}{1 - e^{-\frac{2}{\varepsilon}}} \right),$$

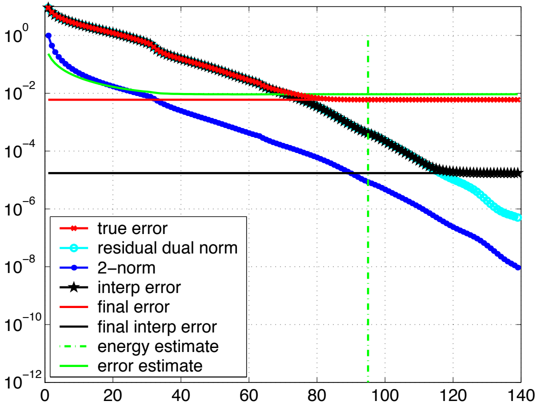
$$\|v_h\|^2_{\mathfrak{H}_h} = \varepsilon |v_h|^2_1 + \sum_{T \in \mathcal{T}^h} \delta_T \|\mathbf{b} \cdot \nabla v_h\|^2_{0,T}$$

Advection-diffusion problem



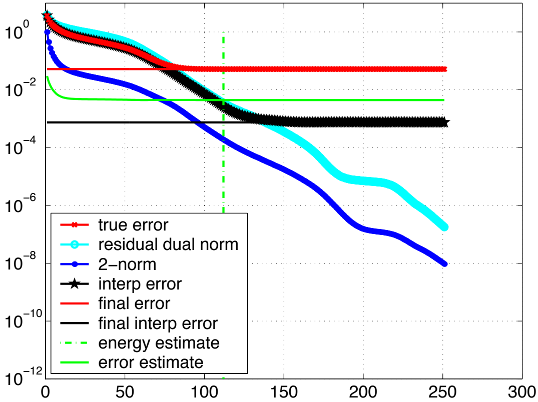
$\varepsilon = 10^{-2}$

Advection-diffusion problem



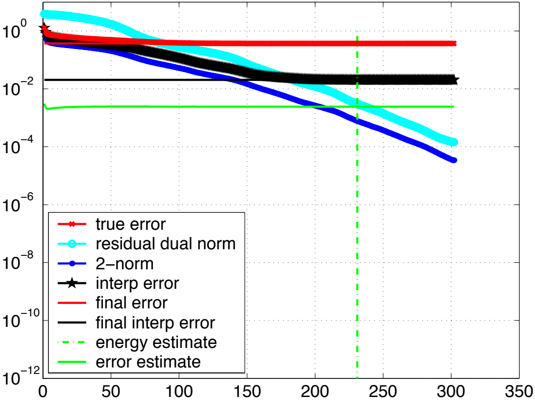
Uniform mesh; $\varepsilon = 1$

Advection-diffusion problem



Uniform mesh; $\varepsilon = 10^{-1}$

Advection-diffusion problem



Uniform mesh; $\varepsilon = 10^{-2}$

How to calculate $\|\mathbf{r}^k\|_{\mathbf{H}^{-1}}$?

- Solve preconditioned system

$$\mathbf{H}^{-1/2}\mathbf{A}\mathbf{H}^{-1/2}\hat{\mathbf{u}} = \mathbf{H}^{-1/2}\mathbf{f}$$

  - $\|\hat{\mathbf{r}}^k\|_{l_2} = \|\mathbf{r}^k\|_{\mathbf{H}^{-1}}$
  - 3-term recurrence.

How to calculate $\|\mathbf{r}^k\|_{\mathbf{H}^{-1}}$?

- Concus & Golub, Widlund: 3-term recurrences for nonsymmetric problems
  - work in $\mathbf{H}$-inner product
  - do not minimize the residual norm.

Recall

$$\mathcal{K}_k(\mathbf{r}^0, \mathbf{A}) = \text{span}\left\{\mathbf{r}^0, \mathbf{A}\mathbf{r}^0, \ldots, \mathbf{A}^{k-1}\mathbf{r}^0\right\}$$

Arnoldi process

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

where $\mathbf{V}_k^T \mathbf{V}_k = \mathbf{I}_k$ and $\mathbf{H}_k$= Hessenberg.

How to calculate $\|\mathbf{r}^k\|_{\mathbf{H}^{-1}}$?

Lemma *Arnoldi applied to*

$$\mathcal{K}_k(\hat{\mathbf{r}}^0, \hat{\mathbf{A}}) \equiv \mathcal{K}_k(\mathbf{H}^{-1/2}\mathbf{r}^0, \mathbf{H}^{-1/2}\mathbf{A}\mathbf{H}^{-1/2})$$

and Arnoldi in the **H***-inner product applied to*

$$\mathcal{K}_k(\tilde{\mathbf{r}}^0, \tilde{A}) \equiv \mathcal{K}_k(\mathbf{H}^{-1}\mathbf{r}^0, \mathbf{H}^{-1}\mathbf{A})$$

produce the same $\mathbf{H}_k$. Moreover,

$$(\mathbf{H}_k)_{ij} = 0, \qquad |i-j| > 1.$$

CG Conclusions

FINAL MESSAGE: DO NOT ACCURATELY COMPUTE THE SOLUTION OF AN INACCURATE PROBLEM

Linear operators

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 $\mathbf{A} \in \mathrm{R}^{m \times n}$ be a full rank matrix.

$$\mathfrak{M} = \{\mathbf{v} \in \mathrm{R}^m; \|\mathbf{u}\|^2_{\mathbf{M}} = \mathbf{v}^T\mathbf{M}\mathbf{v}\},\ \mathfrak{N} = \{\mathbf{q} \in \mathrm{R}^n; \|\mathbf{q}\|^2_{\mathbf{N}} = \mathbf{q}^T\mathbf{N}\mathbf{q}\}$$

$$\mathfrak{M}^\star = \{\mathbf{w} \in \mathrm{R}^m; \|\mathbf{w}\|^2_{\mathbf{M}^{-1}} = \mathbf{w}^T\mathbf{M}^{-1}\mathbf{w}\},$$
$$\mathfrak{N}^\star = \{\mathbf{y} \in \mathrm{R}^n; \|\mathbf{y}\|^2_{\mathbf{N}^{-1}} = \mathbf{y}^T\mathbf{N}^{-1}\mathbf{y}\}$$

Linear operators

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 $\mathbf{A} \in \mathrm{R}^{m\times n}$ be a full rank matrix.

$$\mathfrak{M} = \{\mathbf{v} \in \mathbf{R}^m; \|\mathbf{u}\|^2_{\mathbf{M}} = \mathbf{v}^T\mathbf{M}\mathbf{v}\},\ \mathfrak{N} = \{\mathbf{q} \in \mathbf{R}^n; \|\mathbf{q}\|^2_{\mathbf{N}} = \mathbf{q}^T\mathbf{N}\mathbf{q}\}$$

$$\mathfrak{M}^\star = \{\mathbf{w} \in \mathbf{R}^m; \|\mathbf{w}\|^2_{\mathbf{M}^{-1}} = \mathbf{w}^T\mathbf{M}^{-1}\mathbf{w}\},$$
$$\mathfrak{N}^\star = \{\mathbf{y} \in \mathbf{R}^n; \|\mathbf{y}\|^2_{\mathbf{N}^{-1}} = \mathbf{y}^T\mathbf{N}^{-1}\mathbf{y}\}$$

$$\langle \mathbf{v}, \mathbf{A}\mathbf{q}\rangle_{\mathfrak{M},\mathfrak{M}^\star} = \mathbf{v}^T\mathbf{A}\mathbf{q}, \quad \mathbf{A}\mathbf{q} \in \mathcal{L}(\mathfrak{M})\ \forall \mathbf{q} \in \mathfrak{N}.$$

Linear operators

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 $\mathbf{A} \in \mathrm{R}^{m\times n}$ be a full rank matrix.

$$\mathfrak{M} = \{\mathbf{v} \in \mathrm{R}^m; \|\mathbf{u}\|^2_{\mathbf{M}} = \mathbf{v}^T\mathbf{M}\mathbf{v}\},\ \mathfrak{N} = \{\mathbf{q} \in \mathrm{R}^n; \|\mathbf{q}\|^2_{\mathbf{N}} = \mathbf{q}^T\mathbf{N}\mathbf{q}\}$$

$$\mathfrak{M}^\star = \{\mathbf{w} \in \mathrm{R}^m; \|\mathbf{w}\|^2_{\mathbf{M}^{-1}} = \mathbf{w}^T\mathbf{M}^{-1}\mathbf{w}\},$$
$$\mathfrak{N}^\star = \{\mathbf{y} \in \mathrm{R}^n; \|\mathbf{y}\|^2_{\mathbf{N}^{-1}} = \mathbf{y}^T\mathbf{N}^{-1}\mathbf{y}\}$$

$$\langle \mathbf{v}, \mathbf{A}\mathbf{q}\rangle_{\mathfrak{M},\mathfrak{M}^\star} = \mathbf{v}^T\mathbf{A}\mathbf{q}, \quad \mathbf{A}\mathbf{q} \in \mathcal{L}(\mathfrak{M})\ \forall \mathbf{q} \in \mathfrak{N}.$$

The adjoint operator $\mathbf{A}^\bigstar$ of $\mathbf{A}$ can be defined as

$$\langle \mathbf{A}^\bigstar\mathbf{g}, \mathbf{f}\rangle_{\mathfrak{N}^\star,\mathfrak{N}} = \mathbf{f}^T\mathbf{A}^T\mathbf{g}, \quad \mathbf{A}^T\mathbf{g} \in \mathcal{L}(\mathfrak{N})\ \forall \mathbf{g} \in \mathfrak{M}.$$

Elliptic SVD

Given $\mathbf{q} \in \mathfrak{M}$ and $\mathbf{v} \in \mathfrak{N}$, the critical points for the functional

$$\frac{\mathbf{v}^T \mathbf{A} \mathbf{q}}{\|\mathbf{q}\|_{\mathbf{N}} \ \|\mathbf{v}\|_{\mathbf{M}}}$$

are the "*elliptic singular values and singular vectors*" of $\mathbf{A}$.

Elliptic SVD

Given $\mathbf{q} \in \mathfrak{M}$ and $\mathbf{v} \in \mathfrak{N}$, the critical points for the functional

$$\frac{\mathbf{v}^T \mathbf{A} \mathbf{q}}{\|\mathbf{q}\|_{\mathbf{N}} \, \|\mathbf{v}\|_{\mathbf{M}}}$$

are the "*elliptic singular values and singular vectors*" of **A**.
The saddle-point conditions are

$$\left\{ \begin{array}{lcl} \mathbf{A}\mathbf{q}_i & = & \sigma_i \mathbf{M} \mathbf{v}_i \\ \mathbf{A}^T \mathbf{v}_i & = & \sigma_i \mathbf{N} \mathbf{q}_i \end{array} \right. \qquad \begin{array}{l} \mathbf{v}_i^T \mathbf{M} \mathbf{v}_j = \delta_{ij} \\ \mathbf{q}_i^T \mathbf{N} \mathbf{q}_j = \delta_{ij} \end{array}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$$

Elliptic SVD

Given $\mathbf{q} \in \mathfrak{M}$ and $\mathbf{v} \in \mathfrak{N}$, the critical points for the functional

$$\frac{\mathbf{v}^T \mathbf{A}\mathbf{q}}{\|\mathbf{q}\|_{\mathbf{N}} \, \|\mathbf{v}\|_{\mathbf{M}}}$$

are the "*elliptic singular values and singular vectors*" of **A**.
The saddle-point conditions are

$$\begin{cases} \mathbf{A}\mathbf{q}_i &= \sigma_i \mathbf{M}\mathbf{v}_i \qquad \mathbf{v}_i^T \mathbf{M}\mathbf{v}_j = \delta_{ij} \\ \mathbf{A}^T\mathbf{v}_i &= \sigma_i \mathbf{N}\mathbf{q}_i \qquad \mathbf{q}_i^T \mathbf{N}\mathbf{q}_j = \delta_{ij} \end{cases}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$$

The elliptic singular values are the standard singular values of $\tilde{\mathbf{A}} = \mathbf{M}^{-1/2}\mathbf{A}\mathbf{N}^{-1/2}$. The elliptic singular vectors $\mathbf{q}_i$ and $\mathbf{v}_i$, $i = 1, \ldots, n$ are the transformation by $\mathbf{M}^{-1/2}$ and $\mathbf{N}^{-1/2}$ respectively of the left and right standard singular vector of $\tilde{\mathbf{A}}$.

Quadratic programming

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

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

as a projection problem

$$\min_{\mathbf{A}^T\mathbf{u}=\mathbf{b}} \|\mathbf{u}\|^2_{\mathbf{M}}$$

If $\mathbf{W}$ is non singular then we can choose $\nu = 0$.

Augmented system

The augmented system that gives the optimality conditions for the projection problem:

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

Generalized Golub-Kahan bidiagonalization

In Golub Kahan (1965), Paige Saunders (1982), several algorithms for the bidiagonalization of a $m \times n$ matrix are presented. All of them can be theoretically applied to $\tilde{\mathbf{A}}$ and their generalization to $\mathbf{A}$ is straightforward as shown by Bembow (1999). Here, we want specifically to analyse one of the variants known as the "Craig"-variant (see Paige Saunders (1982), Saunders (1995,1997)).

Generalized Golub-Kahan bidiagonalization

$$\left\{\begin{array}{lcl} \mathbf{A}\tilde{\mathbf{Q}} & = & \mathbf{M}\tilde{\mathbf{V}}\begin{bmatrix}\tilde{\mathbf{B}}\\ 0\end{bmatrix} \\ \mathbf{A}^T\tilde{\mathbf{V}} & = & \mathbf{N}\tilde{\mathbf{Q}}\left[\tilde{\mathbf{B}}^T; 0\right] \end{array}\right. \qquad \begin{array}{l} \tilde{\mathbf{V}}^T\mathbf{M}\tilde{\mathbf{V}} = \mathbf{I}_m \\ \tilde{\mathbf{Q}}^T\mathbf{N}\tilde{\mathbf{Q}} = \mathbf{I}_n \end{array}$$

where

$$\tilde{\mathbf{B}} = \begin{bmatrix} \tilde{\alpha}_1 & 0 & 0 & \cdots & 0 \\ \tilde{\beta}_2 & \tilde{\alpha}_2 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & \tilde{\beta}_{n-1} & \tilde{\alpha}_{n-1} & 0 \\ 0 & \cdots & 0 & \tilde{\beta}_n & \tilde{\alpha}_n \\ 0 & \cdots & 0 & 0 & \tilde{\beta}_{n+1} \end{bmatrix}.$$

Generalized Golub-Kahan bidiagonalization

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

where

$$\mathbf{B} = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \beta_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & \alpha_{n-1} & \beta_{n-1} \\ 0 & \cdots & 0 & 0 & \alpha_n \end{bmatrix}.$$

Algorithm

The augmented system that gives the optimality conditions for $\min_{\mathbf{A}^T\mathbf{u}=\mathbf{b}} \|\mathbf{u}\|^2_{\mathbf{M}}$

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

can be transformed by the change of variables

$$\begin{cases} \mathbf{u} = \mathbf{V}\mathbf{z} \\ \mathbf{p} = \mathbf{Q}\mathbf{y} \end{cases}$$

Algorithm

$$\begin{bmatrix} \mathbf{I}_n & 0 & \mathbf{B} \\ 0 & \mathbf{I}_{m-n} & 0 \\ \mathbf{B}^T & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{Q}^T\mathbf{b} \end{bmatrix}.$$

Algorithm

$$\begin{bmatrix} \mathbf{I}_n & \mathbf{B} \\ \mathbf{B}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{Q}^T\mathbf{b} \end{bmatrix}.$$

Algorithm

$$\begin{bmatrix} \mathbf{I}_n & \mathbf{B} \\ \mathbf{B}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{Q}^T\mathbf{b} \end{bmatrix}.$$

$$\mathbf{Q}^T\mathbf{b} = \mathbf{e}_1\|\mathbf{b}\|_{\mathbf{N}}$$

the value of $\mathbf{z}_1$ will correspond to the first column of the inverse of $\mathbf{B}$ multiplied by $\|\mathbf{b}\|_{\mathbf{N}}$.

Algorithm

Thus, we can compute the first column of **B** and of **V**:
$\alpha_1 \mathbf{M}\mathbf{v}_1 = \mathbf{A}\mathbf{q}_1$, such as

$$\begin{aligned}
&\mathbf{w} = \mathbf{M}^{-1}\mathbf{A}\mathbf{q}_1\\
&\alpha_1 = \mathbf{w}^T\mathbf{M}\mathbf{w} = \mathbf{w}\mathbf{A}\mathbf{q}_1\\
&\mathbf{v}_1 = \mathbf{w}/\sqrt{\alpha_1}.
\end{aligned}$$

Algorithm

Thus, we can compute the first column of **B** and of **V**: $\alpha_1 \mathbf{M}\mathbf{v}_1 = \mathbf{A}\mathbf{q}_1$, such as

$$\begin{aligned}
&\mathbf{w} = \mathbf{M}^{-1}\mathbf{A}\mathbf{q}_1\\
&\alpha_1 = \mathbf{w}^T\mathbf{M}\mathbf{w} = \mathbf{w}\mathbf{A}\mathbf{q}_1\\
&\mathbf{v}_1 = \mathbf{w}/\sqrt{\alpha_1}.
\end{aligned}$$

Finally, knowing $\mathbf{q}_1$ and $\mathbf{v}_1$ we can start the recursive relations

$$\begin{aligned}
&\mathbf{g}_{i+1} = \mathbf{N}^{-1}\left(\mathbf{A}^T\mathbf{v}_i - \alpha_i \mathbf{N}\mathbf{q}_i\right)\\
&\beta_{i+1} = \mathbf{g}^T\mathbf{N}\mathbf{g}\\
&\mathbf{q}_{i+1} = \mathbf{g}\ \sqrt{\beta_{i+1}}\\
&\mathbf{w} = \mathbf{M}^{-1}\left(\mathbf{A}\mathbf{q}_{i+1} - \beta_{i+1}\mathbf{M}\mathbf{v}_i\right)\\
&\alpha_{i+1} = \mathbf{w}^T\mathbf{M}\mathbf{w}\\
&\mathbf{v}_{i+1} = \mathbf{w}/\sqrt{\alpha_{i+1}}.
\end{aligned}$$

$\mathbf{u}$

Thus, the value of $\mathbf{u}$ can be approximated when we have computed the first k columns of $\mathbf{V}$ by

$$\mathbf{u}^{(k)} = \mathbf{V}_k \mathbf{z}_k = \sum_{j=1}^{k} \zeta_j \mathbf{v}_j .$$

$\mathbf{u}$

Thus, the value of **u** can be approximated when we have computed the first k columns of $\mathbf{V}$ by

$$\mathbf{u}^{(k)} = \mathbf{V}_k \mathbf{z}_k = \sum_{j=1}^{k} \zeta_j \mathbf{v}_j.$$

The entries ζ_j of $\mathbf{z}_k$ can be easily computed recursively starting with

$$\zeta_1 = -\frac{\|\mathbf{b}\|_{\mathbf{N}}}{\alpha_1}$$

as

$$\zeta_{i+1} = -\frac{\beta_i}{\alpha_{i+1}}\zeta_i \qquad i = 1, \ldots, n$$

p

Approximating $\mathbf{p} = \mathbf{Q}\mathbf{y}$ by $\mathbf{p}^{(k)} = \mathbf{Q}_k\mathbf{y}_k = \sum_{j=1}^{k} \psi_j \mathbf{q}_j$, we have that

$$\mathbf{y}_k = -\mathbf{B}_k^{-1}\mathbf{z}_k.$$

p

Approximating $\mathbf{p} = \mathbf{Q}\mathbf{y}$ by $\mathbf{p}^{(k)} = \mathbf{Q}_k\mathbf{y}_k = \sum_{j=1}^{k} \psi_j \mathbf{q}_j$, we have that

$$\mathbf{y}_k = -\mathbf{B}_k^{-1}\mathbf{z}_k.$$

Following an observation made by Paige and Saunders, we can easily transform the previous relation into a recursive one where only one extra vector is required.

p

Approximating $\mathbf{p} = \mathbf{Q}\mathbf{y}$ by $\mathbf{p}^{(k)} = \mathbf{Q}_k\mathbf{y}_k = \sum_{j=1}^{k} \psi_j \mathbf{q}_j$, we have that

$$\mathbf{y}_k = -\mathbf{B}_k^{-1}\mathbf{z}_k.$$

From $\mathbf{p}^{(k)} = -\mathbf{Q}_k\mathbf{B}_k^{-1}\mathbf{z}_k = -\left(\mathbf{B}_k^{-T}\mathbf{Q}_k^{T}\right)^{T}\mathbf{z}_k$ and $\mathbf{D}_k = \mathbf{B}_k^{-T}\mathbf{Q}_k^{T}$

$$\mathbf{d}_i = \frac{\mathbf{q}_i - \beta_i\mathbf{d}_{i-1}}{\alpha_i} \qquad i = 1, \ldots, n \;\left(\mathbf{d}_0 = 0\right)$$

where $\mathbf{d}_j$ are the columns of $\mathbf{D}$.
Starting with $\mathbf{p}^{(1)} = -\zeta_1\mathbf{d}_1$ and $\mathbf{u}^{(1)} = \zeta_1\mathbf{v}_1$

$$\left.\begin{array}{l} \mathbf{u}^{(i+1)} = \mathbf{u}^{(i)} + \zeta_{i+1}\mathbf{v}_{i+1} \\ \mathbf{p}^{(i+1)} = \mathbf{p}^{(i)} - \zeta_{i+1}\mathbf{d}_{i+1} \end{array}\right\} \qquad i = 1, \ldots, n$$

Stopping criteria

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

Stopping criteria

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

$$\|\mathbf{A}^T \mathbf{u}^{(k)} - \mathbf{b}\|_{\mathbf{N}^{-1}} = |\beta_{k+1}\, \zeta_k| \leq \sigma_1 |\zeta_k| = \|\tilde{\mathbf{A}}\|_2 |\zeta_k| .$$

Stopping criteria

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

$$\|\mathbf{A}^T \mathbf{u}^{(k)} - \mathbf{b}\|_{\mathbf{N}^{-1}} = |\beta_{k+1}\, \zeta_k| \le \sigma_1 |\zeta_k| = \|\tilde{\mathbf{A}}\|_2 |\zeta_k| .$$

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

Error bound

Lower bound We can estimate $\|\mathbf{e}^{(k)}\|^2_{\mathbf{M}}$ by the lower bound

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

Given a threshold $\tau < 1$ and an integer d, we can stop the iterations when

$$\xi^2_{k,d} \leq \tau \sum_{j=1}^{k+d+1} \zeta_j^2 < \tau \|\mathbf{u}\|^2_{\mathbf{M}}.$$

Error bound

Lower bound We can estimate $\|\mathbf{e}^{(k)}\|^2_{\mathbf{M}}$ by the lower bound

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

Given a threshold $\tau < 1$ and an integer d, we can stop the iterations when

$$\xi^2_{k,d} \leq \tau \sum_{j=1}^{k+d+1} \zeta_j^2 < \tau \|\mathbf{u}\|^2_{\mathbf{M}}.$$

Upper bound Despite being very inexpensive, the previous estimator is still a lower bound of the error. We can use an approach inspired by the Gauss-Radau quadrature algorithm and similar to the one described in Golub-Meurant (2010).

inf-sup

Let $\mathfrak{H}$ and $\mathfrak{P}$ be two Hilbert spaces, and $\mathfrak{H}^\star$ and $\mathfrak{P}^\star$ the corresponding dual spaces. Let

$$\begin{array}{l} \mathfrak{a}(u,v) : \mathfrak{H} \times \mathfrak{H} \to \mathrm{R} \qquad \mathfrak{b}(u,q) : \mathfrak{H} \times \mathfrak{P} \to \mathrm{R} \\ |\mathfrak{a}(u,v)| \leq \|\mathfrak{a}\| \; \|u\|_{\mathfrak{H}} \; \|u\|_{\mathfrak{H}} \quad \forall u \in \mathfrak{H}, \forall v \in \mathfrak{H} \\ |\mathfrak{b}(u,q)| \leq \|\mathfrak{b}\| \; \|v\|_{\mathfrak{H}} \; \|q\|_{\mathfrak{P}} \quad \forall u \in \mathfrak{H}, \forall q \in \mathfrak{P} \end{array}$$

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

$$\begin{array}{ll} \mathfrak{a}(u,v) + \mathfrak{b}(v,p) & = \langle f, v \rangle_{\mathfrak{H}^\star,\mathfrak{H}} \quad \forall v \in \mathfrak{H} \\ \mathfrak{b}(u,q) & = \langle g, q \rangle_{\mathfrak{P}^\star,\mathfrak{P}} \quad \forall q \in \mathfrak{P}. \end{array} \tag{2}$$

inf-sup

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

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

and we have

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

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

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

However, Brezzi:1991 the coercivity on the kernel of $\mathscr{A}^\bigstar$, $Ker(\mathscr{A}^\bigstar)$ is sufficient. We will also assume that $\exists \chi_0 > 0$ such that

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

inf-sup

Under these hypotheses, and for any $f \in \mathfrak{H}^\star$ and $g \in Im(\mathscr{A}^\bigstar)$ then there exists (u, p) solution of saddle problem: u is unique and p is definite up to an element of $Ker(\mathscr{A})$.

inf-sup and Mixed finite-element method

Let now $\mathfrak{H}_h \hookrightarrow \mathfrak{H}$ and $\mathfrak{P}_h \hookrightarrow \mathfrak{P}$ be two finite dimensional subspaces of $\mathfrak{H}$ and $\mathfrak{P}$. As for the problem (2), we can introduce the operators $\mathscr{A}_h : \mathfrak{P}_h \to \mathfrak{H}_h^\star$ and $\mathscr{M}_h ; \mathfrak{H}_h \to \mathfrak{H}_h^\star$. We also assume that

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

inf-sup and Mixed finite-element method

Under the hypotheses of inf-sup, we have that $\exists (u_h, p_h) \in \mathfrak{H}_h \times \mathfrak{P}_h$ solution of

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

and

$$\begin{aligned} \|u - u_h\|_{\mathfrak{H}} \quad + \quad & \|p - p_h\|_{\mathfrak{P} \backslash Ker(A)} \leq \\ & \kappa \left(\inf_{v_h \in \mathfrak{H}_h} \|u - v_h\|_{\mathfrak{H}} + \inf_{q_h \in \mathfrak{P}_h} \|p - q_h\|_{\mathfrak{P}} \right), \end{aligned}$$

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

inf-sup and Mixed finite-element method

Let $\{\phi_i\}_{i=1,\dots,m}$ be a basis for $\mathfrak{H}_h$ and $\{\psi_j\}_{j=1,\dots,n}$ be a basis for $\mathfrak{P}_h$. Then, the matrices **M** and **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 **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 $\mathbf{Aq}_1 \neq 0$, it will compute a matrix **B** of rank less than or equal to $n - s$.

inf-sup and Mixed finite-element method

On the basis of the latter observations, the error $\|\mathbf{e}^{(k)}\|_{\mathbf{M}}$ can be still computed. Finally, we point out 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.

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

inf-sup and Mixed finite-element method

Theorem

Under previous hypotheses, and denoting by $\mathbf{u}^*$ one of the iterates of Algorithm Craig for which $\|\mathbf{e}^{(k)}\|_{\mathbf{M}} < \tau$, we have

$$\begin{aligned}\|u - u^*\|_{\mathfrak{H}} \quad + \quad & \|p - p^*\|_{\mathfrak{P}\setminus Ker(\mathscr{A})} \leq \\ & \check{\kappa}\left(\inf_{v_h \in \mathfrak{H}_h} \|u - v_h\|_{\mathfrak{H}} + \inf_{q_h \in \mathfrak{P}_h} \|p - q_h\|_{\mathfrak{P}} + \tau\right), \end{aligned} \qquad (2)$$

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

Two examples

Stokes

The Stokes problems have been generated using the software provided by **ifiss3.0** package (Elman, Ramage, and Silvester). We use the default geometry of "Step case" and the **Q2-Q1** approximation described in **ifiss3.0** manual and in Elman, Silvester, and Wathen (2005).

name	m	n	nnz($\mathbf{M}$)	nnz($\mathbf{A}$)
Step1	418	61	2126	1603
Step2	1538	209	10190	7140
Step3	5890	769	44236	30483
Step4	23042	2945	184158	126799
Step5	91138	11521	751256	518897

(nnz($\mathbf{M}$) is only for the symmetric part)

Two examples

name	# Iter.s	$\|\mathbf{e}^{(k)}\|_2$	$\|\mathbf{A}^T\mathbf{u}^{(k)} - \mathbf{b}\|_2$	$\|\mathbf{p} - \mathbf{p}^{(k)}\|_2$	$\kappa(\mathbf{B})$
Step1	30	6.8e-16	5.1e-16	1.1e-13	7.6
Step2	32	5.4e-14	5.4e-14	5.0e-12	7.7
Step3	34	3.8e-14	2.7e-14	1.0e-11	7.8
Step4	34	5.0e-13	1.3e-13	1.4e-10	7.8
Step5	35	1.8e-13	3.1e-14	1.7e-10	7.8

Stokes (Step) problems results ($d = 5$, $\tau = 10^{-8}$).

Two examples

Poisson with mixed b.c. Problems The Poisson problem is casted in its dual form as a Darcy's problem:

$$\left\{ \begin{array}{l} \text{Find} \quad w \in \mathfrak{H} = \{\vec{q} \mid \vec{q} \in H_{div}(\Omega),\ \vec{q}\cdot\mathbf{n} = 0 \ \text{ on } \partial_N(\Omega)\},\ u \in L^2(\Omega) \\ \int_\Omega \vec{w}\cdot\vec{q} + \int_{\Omega)} div(\vec{q})u = \int_{\partial_D(\Omega)} u_D\vec{q}\cdot\mathbf{n} \ \ \forall \vec{q} \in \mathfrak{H} \\ \int_\Omega div(\vec{w})v = \int_\Omega fv \ \ \forall v \in L^2(\Omega). \end{array} \right.$$

We approximated the spaces $\mathfrak{H}$ and $L^2(\Omega)$ by RT0 and by piecewise constant functions respectively The matrix $\mathbf{N}$ is the mass matrix for the piecewise constant functions and it is a diagonal matrix with diagonal entries equal to the area of the corresponding triangle. The matrix $\mathbf{M}$ has been chosen such that each approximation $\mathfrak{H}_h$ of $\mathfrak{H}$ is

$$\mathfrak{H}_h = \left\{ \mathbf{q} \in \mathrm{R}^m \ \|\mathbf{q}\|^2_{\mathfrak{H}_h} = \mathbf{q}^T\mathbf{M}\mathbf{q} \right\}.$$

Therefore, denoting by $\mathbf{W}$ the mass matrix for $\mathfrak{H}_h$, we have

$$\mathbf{M} = \mathbf{W} + \mathbf{A}\mathbf{N}^{-1}\mathbf{A}^T.$$

Two examples

Poisson with mixed b.c. Problems

$h = 2^{-k}$	m	n	nnz($\mathbf{M}$)	nnz($\mathbf{A}$)
2^{-6}	12288	8192	36608	24448
2^{-7}	49152	32768	146944	98048
2^{-8}	196608	131072	588800	392704
2^{-9}	786432	524288	2357248	1571840

(nnz($\mathbf{M}$) is only for the symmetric part)

With the chosen boundary conditions, it is easy to verify that the continuous solution u is $u(x, y) = x$.
We point out that the pattern of $\mathbf{W}$ is structurally equal to the pattern $\mathbf{A}\mathbf{N}^{-1}\mathbf{A}^T$.

Two examples

name	# Iter.s	$\|\mathbf{e}^{(k)}\|_2$	$\|\mathbf{A}^T\mathbf{u}^{(k)} - \mathbf{b}\|_2$	$\|\mathbf{p} - \mathbf{p}^{(k)}\|_2$	$\kappa(\mathbf{B})$
$h = 2^{-6}$	10	2.8e-12	2.9e-16	4.1e-11	1.05
$h = 2^{-7}$	10	9.7e-12	3.0e-16	2.6e-10	1.05
$h = 2^{-8}$	10	2.5e-11	3.0e-16	7.9e-10	1.05
$h = 2^{-9}$	10	2.9e-10	2.8e-16	1.3e-08	1.05

Poisson with mixed b.c. data and RT0 problem results ($d = 5$, $\tau = 10^{-8}$).

Lecture on inf-sup

inf-sup

Let $\mathfrak{H}_1$ and $\mathfrak{H}_2$ be two Hilbert space, and $\mathfrak{H}_1^\star$ and $\mathfrak{H}_2^\star$ the corresponding dual spaces. Let

$$\begin{array}{l} \mathfrak{a}(u,v) : \mathfrak{H}_1 \times \mathfrak{H}_2 \to \mathsf{R} \\ \sup_{u\in\mathfrak{H}_1} \sup_{v\in\mathfrak{H}_2} \dfrac{|\mathfrak{a}(u,v)|}{\|u\|_{\mathfrak{H}_1}\ \|v\|_{\mathfrak{H}_2}} \le C_1 \quad \forall u, \in \mathfrak{H}_1, \forall v \in \mathfrak{H}_2 \\ \inf_{u\in\mathfrak{H}_1} \sup_{v\in\mathfrak{H}_2} \dfrac{|\mathfrak{a}(u,v)|}{\|u\|_{\mathfrak{H}_1}\ \|v\|_{\mathfrak{H}_2}} \ge C_2 \quad \forall u \in \mathfrak{H}_1, \forall v \in \mathfrak{H}_2 \end{array}$$

be continuous bilinear forms with $\|\mathfrak{a}\|$ the corresponding norms. Given $f \in \mathfrak{H}_2^\star$, we seek the solutions $u \in \mathfrak{H}_1$ of

$$\mathfrak{a}(u,v) = \langle f, v\rangle_{\mathfrak{H}_2^\star,\mathfrak{H}_2} \quad \forall v \in \mathfrak{H}_2 \tag{3}$$

inf-sup

Theorem. The inf-sup condition is equivalent to

$$\forall v \in \mathfrak{H}_2 \exists u \in \mathfrak{H}_1 \text{ s.t.}$$
$$\mathfrak{a}(u,v) \geq c_1 ||v||^2_{\mathfrak{H}_2} \text{ and } ||u||_{\mathfrak{H}_1} \leq c_2 ||v||_{\mathfrak{H}_2}.$$

IF $\mathfrak{H}_1 = \mathfrak{H}_2$ THEN the inf-sup is the coercivity condition

inf-sup

Solve (we assume that we have approximate the Hilbert spaces with finite dimensional ones)

$$\mathbf{A}\mathbf{u} = \mathbf{f}$$

given

$$\begin{aligned} \max_{\mathbf{w}\in\mathrm{R}^n\setminus\{\mathbf{0}\}} \max_{\mathbf{v}\in\mathrm{R}^n\setminus\{\mathbf{0}\}} \frac{\mathbf{w}^T\mathbf{A}\mathbf{v}}{\|\mathbf{v}\|_{\mathbf{H}}\|\mathbf{w}\|_{\mathbf{H}}} &\leq C_1 && \text{(sup-sup)} \\ \min_{\mathbf{w}\in\mathrm{R}^n\setminus\{\mathbf{0}\}} \max_{\mathbf{v}\in\mathrm{R}^n\setminus\{\mathbf{0}\}} \frac{\mathbf{w}^T\mathbf{A}\mathbf{v}}{\|\mathbf{v}\|_{\mathbf{H}}\|\mathbf{w}\|_{\mathbf{H}}} &\geq C_2 && \text{(inf-sup)} \end{aligned}$$

Note: $\|v_h\|_{\mathfrak{H}_h} = \|\mathbf{v}\|_{\mathbf{H}}$ defines the spd matrix $\mathbf{H}$.

inf-sup

Let $\mathfrak{H}$ and $\mathfrak{P}$ be two Hilbert spaces, and $\mathfrak{H}^\star$ and $\mathfrak{P}^\star$ the corresponding dual spaces. Let

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

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

$$\begin{array}{lll}
\mathfrak{a}(u,v) + \mathfrak{b}(v,p) & = \langle f, v \rangle_{\mathfrak{H}^\star,\mathfrak{H}} & \forall v \in \mathfrak{H} \\
\mathfrak{b}(u,q) & = \langle g, q \rangle_{\mathfrak{P}^\star,\mathfrak{P}} & \forall q \in \mathfrak{P}.
\end{array} \tag{3}$$

inf-sup

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

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

and we have

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

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

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

However, Brezzi:1991 the coercivity on the kernel of $\mathscr{A}^\bigstar$, $Ker(\mathscr{A}^\bigstar)$ is sufficient. We will also assume that $\exists \chi_0 > 0$ such that

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

inf-sup

Under these hypotheses, and for any $f \in \mathfrak{H}^\star$ and $g \in Im(\mathscr{A}^\bigstar)$ then there exists (u, p) solution of saddle problem: u is unique and p is definite up to an element of $Ker(\mathscr{A})$.

inf-sup

Remember that

$$\langle \mathscr{A}^{\bigstar} u, q \rangle_{\mathfrak{P}^\star \times \mathfrak{P}} \text{ and } \langle v, \mathscr{A} p \rangle_{\mathfrak{H} \times \mathfrak{H}^\star} = \mathfrak{b}(v, p)$$

Then, we solve

$$\left[\begin{array}{cc} \mathscr{M} & \mathscr{A} \\ \mathscr{A}^{\bigstar} & \end{array} \right] \left[\begin{array}{c} u \\ p \end{array} \right] = \left[\begin{array}{c} f \\ g \end{array} \right]$$

inf-sup and Mixed finite-element method

Let now $\mathfrak{H}_h \hookrightarrow \mathfrak{H}$ and $\mathfrak{P}_h \hookrightarrow \mathfrak{P}$ be two finite dimensional subspaces of $\mathfrak{H}$ and $\mathfrak{P}$. As for the problem (2), we can introduce the operators $\mathscr{A}_h : \mathfrak{P}_h \to \mathfrak{H}_h^\star$ and $\mathscr{M}_h ; \mathfrak{H}_h \to \mathfrak{H}_h^\star$. We also assume that

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

inf-sup and Mixed finite-element method

Under the hypotheses of inf-sup, we have that
$\exists (u_h, p_h) \in \mathfrak{H}_h \times \mathfrak{P}_h$ solution of

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

and

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

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

inf-sup and Mixed finite-element method

Let $\{\phi_i\}_{i=1,\dots,m}$ be a basis for $\mathfrak{H}_h$ and $\{\psi_j\}_{j=1,\dots,n}$ be a basis for $\mathfrak{P}_h$. 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 $n - s$.

inf-sup and Mixed finite-element method

On the basis of the latter observations, the error $\|\mathbf{e}^{(k)}\|_{\mathbf{M}}$ can be still computed. Finally, we point out 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.

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

Generalized SVD

Given $\mathbf{q} \in \mathfrak{M}$ and $\mathbf{v} \in \mathfrak{N}$, the critical points for the functional

$$\frac{\mathbf{v}^T \mathbf{A}\mathbf{q}}{\|\mathbf{q}\|_{\mathbf{N}} \ \|\mathbf{v}\|_{\mathbf{M}}}$$

are the "*elliptic singular values and singular vectors*" of $\mathbf{A}$.

Generalized SVD

Given $\mathbf{q} \in \mathfrak{M}$ and $\mathbf{v} \in \mathfrak{N}$, the critical points for the functional

$$\frac{\mathbf{v}^T \mathbf{A} \mathbf{q}}{\|\mathbf{q}\|_{\mathbf{N}} \, \|\mathbf{v}\|_{\mathbf{M}}}$$

are the "*elliptic singular values and singular vectors*" of **A**.
The saddle-point conditions are

$$\left\{ \begin{array}{rcl} \mathbf{A}\mathbf{q}_i & = & \sigma_i \mathbf{M}\mathbf{v}_i \\ \mathbf{A}^T\mathbf{v}_i & = & \sigma_i \mathbf{N}\mathbf{q}_i \end{array} \right. \qquad \begin{array}{l} \mathbf{v}_i^T \mathbf{M}\mathbf{v}_j = \delta_{ij} \\ \mathbf{q}_i^T \mathbf{N}\mathbf{q}_j = \delta_{ij} \end{array}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$$

Generalized SVD

Given $\mathbf{q} \in \mathfrak{M}$ and $\mathbf{v} \in \mathfrak{N}$, the critical points for the functional

$$\frac{\mathbf{v}^T\mathbf{A}\mathbf{q}}{\|\mathbf{q}\|_{\mathbf{N}}\ \|\mathbf{v}\|_{\mathbf{M}}}$$

are the "*elliptic singular values and singular vectors*" of $\mathbf{A}$.
The saddle-point conditions are

$$\begin{cases} \mathbf{A}\mathbf{q}_i &= \sigma_i \mathbf{M}\mathbf{v}_i \qquad \mathbf{v}_i^T\mathbf{M}\mathbf{v}_j = \delta_{ij} \\ \mathbf{A}^T\mathbf{v}_i &= \sigma_i \mathbf{N}\mathbf{q}_i \qquad \mathbf{q}_i^T\mathbf{N}\mathbf{q}_j = \delta_{ij} \end{cases}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$$

The elliptic singular values are the standard singular values of $\tilde{\mathbf{A}} = \mathbf{M}^{-1/2}\mathbf{A}\mathbf{N}^{-1/2}$. The elliptic singular vectors $\mathbf{q}_i$ and $\mathbf{v}_i$, $i = 1, \dots, n$ are the transformation by $\mathbf{M}^{-1/2}$ and $\mathbf{N}^{-1/2}$ respectively of the left and right standard singular vector of $\tilde{\mathbf{A}}$.

Quadratic programming

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

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

as a projection problem

$$\min_{\mathbf{A}^T\mathbf{u}=\mathbf{b}} \|\mathbf{u}\|^2_{\mathbf{M}}$$

If $\mathbf{W}$ is non singular then we can choose $\nu = 0$.

Augmented system

The augmented system that gives the optimality conditions for the projection problem:

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

Lecture on Golub-Kahan

Elliptic SVD

Given $\mathbf{q} \in \mathfrak{M}$ and $\mathbf{v} \in \mathfrak{N}$, the critical points for the functional

$$\frac{\mathbf{v}^T \mathbf{A}\mathbf{q}}{\|\mathbf{q}\|_{\mathbf{N}} \; \|\mathbf{v}\|_{\mathbf{M}}}$$

are the "*ELLIPTIC singular values and singular vectors*" of $\mathbf{A}$.

Elliptic SVD

Given $\mathbf{q} \in \mathfrak{M}$ and $\mathbf{v} \in \mathfrak{N}$, the critical points for the functional

$$\frac{\mathbf{v}^T \mathbf{A} \mathbf{q}}{\|\mathbf{q}\|_{\mathbf{N}} \, \|\mathbf{v}\|_{\mathbf{M}}}$$

are the "*ELLIPTIC singular values and singular vectors*" of $\mathbf{A}$.
The saddle-point conditions are

$$\left\{ \begin{array}{lcl} \mathbf{A}\mathbf{q}_i & = & \sigma_i \mathbf{M}\mathbf{v}_i \\ \mathbf{A}^T\mathbf{v}_i & = & \sigma_i \mathbf{N}\mathbf{q}_i \end{array} \right. \qquad \begin{array}{l} \mathbf{v}_i^T \mathbf{M}\mathbf{v}_j = \delta_{ij} \\ \mathbf{q}_i^T \mathbf{N}\mathbf{q}_j = \delta_{ij} \end{array}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$$

Elliptic SVD

Given $\mathbf{q} \in \mathfrak{M}$ and $\mathbf{v} \in \mathfrak{N}$, the critical points for the functional

$$\frac{\mathbf{v}^T \mathbf{A} \mathbf{q}}{\|\mathbf{q}\|_{\mathbf{N}} \ \|\mathbf{v}\|_{\mathbf{M}}}$$

are the "*ELLIPTIC singular values and singular vectors*" of **A**. The saddle-point conditions are

$$\begin{cases} \mathbf{A}\mathbf{q}_i & = & \sigma_i \mathbf{M}\mathbf{v}_i \qquad & \mathbf{v}_i^T \mathbf{M}\mathbf{v}_j = \delta_{ij} \\ \mathbf{A}^T\mathbf{v}_i & = & \sigma_i \mathbf{N}\mathbf{q}_i \qquad & \mathbf{q}_i^T \mathbf{N}\mathbf{q}_j = \delta_{ij} \end{cases}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$$

The elliptic singular values are the standard singular values of $\tilde{\mathbf{A}} = \mathbf{M}^{-1/2}\mathbf{A}\mathbf{N}^{-1/2}$. The elliptic singular vectors $\mathbf{q}_i$ and $\mathbf{v}_i$, $i = 1, \dots, n$ are the transformation by $\mathbf{M}^{-1/2}$ and $\mathbf{N}^{-1/2}$ respectively of the left and right standard singular vector of $\tilde{\mathbf{A}}$.

Quadratic programming

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

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

as a projection problem

$$\min_{\mathbf{A}^T\mathbf{u}=\mathbf{b}} \|\mathbf{u}\|^2_{\mathbf{M}}$$

If $\mathbf{W}$ is non singular then we can choose $\nu = 0$.

Augmented system

The augmented system that gives the optimality conditions for the projection problem:

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

Generalized Golub-Kahan bidiagonalization

In Golub Kahan (1965), Paige Saunders (1982), several algorithms for the bidiagonalization of a $m \times n$ matrix are presented. All of them can be theoretically applied to $\tilde{\mathbf{A}}$ and their generalization to $\mathbf{A}$ is straightforward as shown by Bembow (1999). Here, we want specifically to analyse one of the variants known as the "Craig"-variant (see Paige Saunders (1982), Saunders (1995,1997)).

Generalized Golub-Kahan bidiagonalization

$$\left\{\begin{array}{rcl} \mathbf{A}\tilde{\mathbf{Q}} & = & \mathbf{M}\tilde{\mathbf{V}}\begin{bmatrix}\tilde{\mathbf{B}}\\ 0\end{bmatrix} \\ \mathbf{A}^T\tilde{\mathbf{V}} & = & \mathbf{N}\tilde{\mathbf{Q}}\left[\tilde{\mathbf{B}}^T;0\right]\end{array}\right. \qquad \begin{array}{l}\tilde{\mathbf{V}}^T\mathbf{M}\tilde{\mathbf{V}} = \mathbf{I}_m \\ \tilde{\mathbf{Q}}^T\mathbf{N}\tilde{\mathbf{Q}} = \mathbf{I}_n\end{array}$$

where

$$\tilde{\mathbf{B}} = \begin{bmatrix} \tilde{\alpha}_1 & 0 & 0 & \cdots & 0 \\ \tilde{\beta}_2 & \tilde{\alpha}_2 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & \tilde{\beta}_{n-1} & \tilde{\alpha}_{n-1} & 0 \\ 0 & \cdots & 0 & \tilde{\beta}_n & \tilde{\alpha}_n \\ 0 & \cdots & 0 & 0 & \tilde{\beta}_{n+1} \end{bmatrix}.$$

Generalized Golub-Kahan bidiagonalization

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

where

$$\mathbf{B} = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \beta_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & \alpha_{n-1} & \beta_{n-1} \\ 0 & \cdots & 0 & 0 & \alpha_n \end{bmatrix}.$$

Algorithm

The augmented system that gives the optimality conditions for $\min_{\mathbf{A}^T\mathbf{u}=\mathbf{b}} \|\mathbf{u}\|^2_{\mathbf{M}}$

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

can be transformed by the change of variables

$$\begin{cases} \mathbf{u} = \mathbf{V}\mathbf{z} \\ \mathbf{p} = \mathbf{Q}\mathbf{y} \end{cases}$$

Algorithm

$$\begin{bmatrix} \mathbf{I}_n & 0 & \mathbf{B} \\ 0 & \mathbf{I}_{m-n} & 0 \\ \mathbf{B}^T & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{Q}^T\mathbf{b} \end{bmatrix}.$$

Algorithm

$$\begin{bmatrix} \mathbf{I}_n & \mathbf{B} \\ \mathbf{B}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{Q}^T\mathbf{b} \end{bmatrix}.$$

Algorithm

$$\begin{bmatrix} \mathbf{I}_n & \mathbf{B} \\ \mathbf{B}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{Q}^T\mathbf{b} \end{bmatrix}.$$

$$\mathbf{Q}^T\mathbf{b} = \mathbf{e}_1\|\mathbf{b}\|_{\mathbf{N}}$$

the value of $\mathbf{z}_1$ will correspond to the first column of the inverse of $\mathbf{B}$ multiplied by $\|\mathbf{b}\|_{\mathbf{N}}$.

Algorithm

Thus, we can compute the first column of **B** and of **V**:
$\alpha_1 \mathbf{M}\mathbf{v}_1 = \mathbf{A}\mathbf{q}_1$, such as

$$\begin{aligned}
&\mathbf{w} = \mathbf{M}^{-1}\mathbf{A}\mathbf{q}_1\\
&\alpha_1 = \mathbf{w}^T\mathbf{M}\mathbf{w} = \mathbf{w}\mathbf{A}\mathbf{q}_1\\
&\mathbf{v}_1 = \mathbf{w}/\sqrt{\alpha_1}.
\end{aligned}$$

Algorithm

Thus, we can compute the first column of $\mathbf{B}$ and of $\mathbf{V}$: $\alpha_1 \mathbf{M}\mathbf{v}_1 = \mathbf{A}\mathbf{q}_1$, such as

$$\begin{array}{l} \mathbf{w} = \mathbf{M}^{-1}\mathbf{A}\mathbf{q}_1 \\ \alpha_1 = \mathbf{w}^T\mathbf{M}\mathbf{w} = \mathbf{w}\mathbf{A}\mathbf{q}_1 \\ \mathbf{v}_1 = \mathbf{w}/\sqrt{\alpha_1}. \end{array}$$

Finally, knowing $\mathbf{q}_1$ and $\mathbf{v}_1$ we can start the recursive relations

$$\begin{array}{l} \mathbf{g}_{i+1} = \mathbf{N}^{-1}\left(\mathbf{A}^T\mathbf{v}_i - \alpha_i \mathbf{N}\mathbf{q}_i\right) \\ \beta_{i+1} = \mathbf{g}^T\mathbf{N}\mathbf{g} \\ \mathbf{q}_{i+1} = \mathbf{g}\ \sqrt{\beta_{i+1}} \\ \mathbf{w} = \mathbf{M}^{-1}\left(\mathbf{A}\mathbf{q}_{i+1} - \beta_{i+1}\mathbf{M}\mathbf{v}_i\right) \\ \alpha_{i+1} = \mathbf{w}^T\mathbf{M}\mathbf{w} \\ \mathbf{v}_{i+1} = \mathbf{w}/\sqrt{\alpha_{i+1}}. \end{array}$$

u

Thus, the value of **u** can be approximated when we have computed the first *k* columns of **V** by

$$\mathbf{u}^{(k)} = \mathbf{V}_k \mathbf{z}_k = \sum_{j=1}^{k} \zeta_j \mathbf{v}_j .$$

$\mathbf{u}$

Thus, the value of $\mathbf{u}$ can be approximated when we have computed the first k columns of $\mathbf{V}$ by

$$\mathbf{u}^{(k)} = \mathbf{V}_k \mathbf{z}_k = \sum_{j=1}^{k} \zeta_j \mathbf{v}_j.$$

The entries ζ_j of $\mathbf{z}_k$ can be easily computed recursively starting with

$$\zeta_1 = -\frac{\|\mathbf{b}\|_{\mathbf{N}}}{\alpha_1}$$

as

$$\zeta_{i+1} = -\frac{\beta_i}{\alpha_{i+1}} \zeta_i \qquad i = 1, \ldots, n$$

p

Approximating $\mathbf{p} = \mathbf{Q}\mathbf{y}$ by $\mathbf{p}^{(k)} = \mathbf{Q}_k\mathbf{y}_k = \sum_{j=1}^{k} \psi_j \mathbf{q}_j$, we have that

$$\mathbf{y}_k = -\mathbf{B}_k^{-1}\mathbf{z}_k.$$

p

Approximating $\mathbf{p} = \mathbf{Q}\mathbf{y}$ by $\mathbf{p}^{(k)} = \mathbf{Q}_k\mathbf{y}_k = \sum_{j=1}^{k} \psi_j \mathbf{q}_j$, we have that

$$\mathbf{y}_k = -\mathbf{B}_k^{-1}\mathbf{z}_k.$$

Following an observation made by Paige and Saunders, we can easily transform the previous relation into a recursive one where only one extra vector is required.

p

Approximating $\mathbf{p} = \mathbf{Q}\mathbf{y}$ by $\mathbf{p}^{(k)} = \mathbf{Q}_k\mathbf{y}_k = \sum_{j=1}^{k} \psi_j \mathbf{q}_j$, we have that

$$\mathbf{y}_k = -\mathbf{B}_k^{-1}\mathbf{z}_k.$$

From $\mathbf{p}^{(k)} = -\mathbf{Q}_k\mathbf{B}_k^{-1}\mathbf{z}_k = -\left(\mathbf{B}_k^{-T}\mathbf{Q}_k^T\right)^T \mathbf{z}_k$ and $\mathbf{D}_k = \mathbf{B}_k^{-T}\mathbf{Q}_k^T$

$$\mathbf{d}_i = \frac{\mathbf{q}_i - \beta_i \mathbf{d}_{i-1}}{\alpha_i} \qquad i = 1, \ldots, n \ \left(\mathbf{d}_0 = 0\right)$$

where $\mathbf{d}_j$ are the columns of $\mathbf{D}$.
Starting with $\mathbf{p}^{(1)} = -\zeta_1\mathbf{d}_1$ and $\mathbf{u}^{(1)} = \zeta_1\mathbf{v}_1$

$$\left.\begin{array}{l} \mathbf{u}^{(i+1)} = \mathbf{u}^{(i)} + \zeta_{i+1}\mathbf{v}_{i+1} \\ \mathbf{p}^{(i+1)} = \mathbf{p}^{(i)} - \zeta_{i+1}\mathbf{d}_{i+1} \end{array}\right\} \qquad i = 1, \ldots, n$$

Stopping criteria

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

Stopping criteria

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

$$\|\mathbf{A}^T\mathbf{u}^{(k)} - \mathbf{b}\|_{\mathbf{N}^{-1}} = |\beta_{k+1}\,\zeta_k| \leq \sigma_1|\zeta_k| = \|\tilde{\mathbf{A}}\|_2|\zeta_k|.$$

Stopping criteria

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

$$\|\mathbf{A}^T\mathbf{u}^{(k)} - \mathbf{b}\|_{\mathbf{N}^{-1}} = |\beta_{k+1}\, \zeta_k| \leq \sigma_1 |\zeta_k| = \|\tilde{\mathbf{A}}\|_2 |\zeta_k| .$$

$$\|\mathbf{p}-\mathbf{p}^{(k)}\|_{\mathbf{N}} = \left\| \mathbf{Q}\mathbf{B}^{-1} \left(\mathbf{z} - \left[\begin{array}{c} \mathbf{z}_k \\ 0 \end{array} \right] \right) \right\|_{\mathbf{N}} \leq \frac{\|\mathbf{e}^{(k)}\|_{\mathbf{M}}}{\sigma_n} .$$

Error bound

Lower bound We can estimate $\|\mathbf{e}^{(k)}\|^2_{\mathbf{M}}$ by the lower bound

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

Given a threshold $\tau < 1$ and an integer d, we can stop the iterations when

$$\xi^2_{k,d} \leq \tau \sum_{j=1}^{k+d+1} \zeta_j^2 < \tau \|\mathbf{u}\|^2_{\mathbf{M}}.$$

Error bound

Lower bound We can estimate $\|\mathbf{e}^{(k)}\|^2_{\mathbf{M}}$ by the lower bound

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

Given a threshold $\tau < 1$ and an integer d, we can stop the iterations when

$$\xi^2_{k,d} \leq \tau \sum_{j=1}^{k+d+1} \zeta_j^2 < \tau \|\mathbf{u}\|^2_{\mathbf{M}}.$$

Upper bound Despite being very inexpensive, the previous estimator is still a lower bound of the error. We can use an approach inspired by the Gauss-Radau quadrature algorithm and similar to the one described in Golub-Meurant (2010).

Error bound

Theorem

Under previous hypotheses, and denoting by $\mathbf{u}^*$ *one of the iterates of Algorithm Craig for which* $\|\mathbf{e}^{(k)}\|_{\mathbf{M}} < \tau$, *we have*

$$\begin{aligned}\|u - u^*\|_{\mathfrak{H}} \quad + \quad & \|p - p^*\|_{\mathfrak{P}\setminus Ker(\mathscr{A})} \leq \\ & \check{\kappa}\left(\inf_{v_h\in\mathfrak{H}_h}\|u - v_h\|_{\mathfrak{H}} + \inf_{q_h\in\mathfrak{P}_h}\|p - q_h\|_{\mathfrak{P}} + \tau\right), \qquad (3)\end{aligned}$$

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

Lecture on SQD

Symmetric Quasi-Definite Systems

$$\begin{bmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & -\mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} \qquad \text{where} \qquad \mathbf{M} = \mathbf{M}^T \succ 0, \; \mathbf{N} = \mathbf{N}^T \succ 0.$$

- Interior-point methods for LP, QP, NLP, SOCP, SDP, ...
- Regularized/stabilized PDE problems
- Regularized least squares
- How to best take advantage of the structure?

Main Property

Theorem (Vanderbei, 1995)
If $\mathbf{K}$ *is SQD, it is* **strongly factorizable***, i.e., for any permutation matrix* $\mathbf{P}$*, there exists a unit lower triangular* $\mathbf{L}$ *and a diagonal* $\mathbf{D}$ *such that* $\mathbf{P}^T\mathbf{K}\mathbf{P} = \mathbf{L}\mathbf{D}\mathbf{L}^T$*.*

- Cholesky-factorizable
- Used to speed up factorization in regularized least-squares (Saunders) and interior-point methods (Friedlander and O.)
- Stability analysis by Gill, Saunders, Shinnerl (1996).

Centered preconditioning

$$\begin{bmatrix} \mathbf{M}^{-\frac{1}{2}} & \\ & \mathbf{N}^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & -\mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{M}^{-\frac{1}{2}} & \\ & \mathbf{N}^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathbf{M}^{-\frac{1}{2}}\mathbf{f} \\ \mathbf{N}^{-\frac{1}{2}}\mathbf{g} \end{bmatrix}$$

which is equivalent to

$$\overbrace{\begin{bmatrix} \mathbf{I}_m & \mathbf{M}^{-\frac{1}{2}}\mathbf{A}\mathbf{N}^{-\frac{1}{2}} \\ \mathbf{N}^{-\frac{1}{2}}\mathbf{A}^T\mathbf{M}^{-\frac{1}{2}} & -\mathbf{I}_n \end{bmatrix}}^{\widehat{\mathbf{C}}} \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathbf{M}^{-\frac{1}{2}}\mathbf{f} \\ \mathbf{N}^{-\frac{1}{2}}\mathbf{g} \end{bmatrix}$$

Centered preconditioning

$$\begin{bmatrix} \mathbf{M}^{-\frac{1}{2}} & \\ & \mathbf{N}^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & -\mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{M}^{-\frac{1}{2}} & \\ & \mathbf{N}^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathbf{M}^{-\frac{1}{2}}\mathbf{f} \\ \mathbf{N}^{-\frac{1}{2}}\mathbf{g} \end{bmatrix}$$

which is equivalent to

$$\overbrace{\begin{bmatrix} \mathbf{I}_m & \mathbf{M}^{-\frac{1}{2}}\mathbf{A}\mathbf{N}^{-\frac{1}{2}} \\ \mathbf{N}^{-\frac{1}{2}}\mathbf{A}^T\mathbf{M}^{-\frac{1}{2}} & -\mathbf{I}_n \end{bmatrix}}^{\widehat{\mathbf{C}}} \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathbf{M}^{-\frac{1}{2}}\mathbf{f} \\ \mathbf{N}^{-\frac{1}{2}}\mathbf{g} \end{bmatrix}$$

Theorem (Saunders (1995))

Suppose $\tilde{\mathbf{A}} = \mathbf{M}^{-\frac{1}{2}}\mathbf{A}\mathbf{N}^{-\frac{1}{2}}$ has rank $p \leq m$ with nonzero singular values $\sigma_1, \ldots, \sigma_p$. The eigenvalues of $\widehat{\mathbf{C}}$ are $+1$, -1 and $\pm\sqrt{1+\sigma_k}$, $k = 1, \ldots, p$.

Symmetric spectrum and Iterative methods

A symmetric matrix with a symmetric spectrum can be transform preserving the symmetry of the spectrum in a SQD one.
Moreover, Fischer (Theorem 6.9.9 in “Polynomial based iteration methods for symmetric linear systems”) Freund (1983), Freund Golub Nachtigal (1992), and Ramage Silvester Wathen (1995) give different poofs that MINRES and CG perform redundant iterations.

Iterative Methods I

Facts: SQD systems are symmetric, non-singular, square and indefinite.

Iterative Methods I

Facts: SQD systems are symmetric, non-singular, square and indefinite.

- MINRES
- SYMMLQ
- (F)GMRES??
- QMRS????

Iterative Methods I

Facts: SQD systems are symmetric, non-singular, square and indefinite.

- MINRES
- SYMMLQ
- (F)GMRES??
- QMRS????

Fact: ... none exploits the SQD structure and they are doing redundant iterations

Related Problems: an example

$$\begin{bmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & -\mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix}$$

Related Problems: an example

$$\begin{bmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & -\mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix}$$

are the optimality conditions of

$$\min_{\mathbf{y}\in\mathbb{R}^m} \frac{1}{2} \left\| \begin{bmatrix} \mathbf{A} \\ \mathbf{I} \end{bmatrix} \mathbf{y} - \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix} \right\|^2_{E_+^{-1}} \equiv \min_{y\in\mathbb{R}^m} \frac{1}{2} \left\| \begin{bmatrix} \mathbf{M}^{-\frac{1}{2}} & 0 \\ 0 & \mathbf{N}^{\frac{1}{2}} \end{bmatrix} \left(\begin{bmatrix} \mathbf{A} \\ \mathbf{I} \end{bmatrix} \mathbf{y} - \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix} \right) \right\|^2_2$$

Related Problems: an example

$$\begin{bmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & -\mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix}$$

are the optimality conditions of

$$\min_{\mathbf{y}\in\mathrm{R}^m} \frac{1}{2} \left\| \begin{bmatrix} \mathbf{A} \\ \mathbf{I} \end{bmatrix} \mathbf{y} - \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix} \right\|^2_{E_+^{-1}} \equiv \min_{y\in\mathrm{R}^m} \frac{1}{2} \left\| \begin{bmatrix} \mathbf{M}^{-\frac{1}{2}} & 0 \\ 0 & \mathbf{N}^{\frac{1}{2}} \end{bmatrix} \left(\begin{bmatrix} \mathbf{A} \\ \mathbf{I} \end{bmatrix} \mathbf{y} - \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix} \right) \right\|^2_2$$

or of

$$\underset{\mathbf{x},\mathbf{y}}{\text{minimize}} \;\; \frac{1}{2}(\|\mathbf{x}\|^2_{\mathbf{M}} + \|\mathbf{y}\|^2_{\mathbf{N}}) \quad \text{subject to } \mathbf{M}\mathbf{x} + \mathbf{A}\mathbf{y} = \mathbf{b}.$$

Some properties of SQD matrices

Let us denote the Cholesky factors of $\mathbf{M}$ and $\mathbf{N}$ by $\mathbf{R}$ and $\mathbf{U}$ (upper triangular matrices).

$$\mathbf{H} = \begin{bmatrix} \mathbf{M} & \\ & \mathbf{N} \end{bmatrix} = \begin{bmatrix} \mathbf{R}^T\mathbf{R} & \\ & \mathbf{U}^T\mathbf{U} \end{bmatrix} = \widetilde{\mathbf{R}}^T\widetilde{\mathbf{R}}$$

We observe that

$$\mathbf{C} = \begin{bmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & -\mathbf{N} \end{bmatrix} = \begin{bmatrix} \mathbf{R}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{U}^T \end{bmatrix} \begin{bmatrix} \mathbf{I}_m & \tilde{\mathbf{A}} \\ \tilde{\mathbf{A}}^T & -\mathbf{I}_n \end{bmatrix} \begin{bmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{U} \end{bmatrix} = \widetilde{\mathbf{R}}^T\widetilde{\mathbf{C}}\widetilde{\mathbf{R}},$$

Some properties of SQD matrices

Let us denote the Cholesky factors of $\mathbf{M}$ and $\mathbf{N}$ by $\mathbf{R}$ and $\mathbf{U}$ (upper triangular matrices).

$$\mathbf{H} = \begin{bmatrix} \mathbf{M} & \\ & \mathbf{N} \end{bmatrix} = \begin{bmatrix} \mathbf{R}^T\mathbf{R} & \\ & \mathbf{U}^T\mathbf{U} \end{bmatrix} = \widetilde{\mathbf{R}}^T\widetilde{\mathbf{R}}$$

We observe that

$$\mathbf{C} = \begin{bmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & -\mathbf{N} \end{bmatrix} = \begin{bmatrix} \mathbf{R}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{U}^T \end{bmatrix} \begin{bmatrix} \mathbf{I}_m & \tilde{\mathbf{A}} \\ \tilde{\mathbf{A}}^T & -\mathbf{I}_n \end{bmatrix} \begin{bmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{U} \end{bmatrix} = \widetilde{\mathbf{R}}^T\widetilde{\mathbf{C}}\widetilde{\mathbf{R}},$$

$$\mathbf{H}^{-1}\mathbf{C} = \widetilde{\mathbf{R}}^{-1}\widetilde{\mathbf{C}}\widetilde{\mathbf{R}}$$

Some properties of SQD matrices

By direct computation it is easy to prove that

$$\widetilde{\mathbf{C}}^2 = \begin{bmatrix} \mathbf{I}_m + \tilde{\mathbf{A}}\tilde{\mathbf{A}}^T & \\ & \mathbf{I}_n + \tilde{\mathbf{A}}^T\tilde{\mathbf{A}} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{D}}_1 & \\ & \widetilde{\mathbf{D}}_2 \end{bmatrix} = \widetilde{\mathbf{D}}.$$

Some properties of SQD matrices

By direct computation it is easy to prove that

$$\widetilde{\mathbf{C}}^2 = \begin{bmatrix} \mathbf{I}_m + \tilde{\mathbf{A}}\tilde{\mathbf{A}}^T & \\ & \mathbf{I}_n + \tilde{\mathbf{A}}^T\tilde{\mathbf{A}} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{D}}_1 & \\ & \widetilde{\mathbf{D}}_2 \end{bmatrix} = \widetilde{\mathbf{D}}.$$

$$\begin{aligned} \widetilde{\mathbf{C}}^{-1} &= \widetilde{\mathbf{D}}^{-1}\widetilde{\mathbf{C}} = \widetilde{\mathbf{C}}\widetilde{\mathbf{D}}^{-1}; \\ \widetilde{\mathbf{C}}\widetilde{\mathbf{D}} &= \widetilde{\mathbf{C}}^3 = \widetilde{\mathbf{D}}\widetilde{\mathbf{C}}; \\ \mathbf{C}\mathbf{H}^{-1}\mathbf{C} &= \widetilde{\mathbf{R}}^T\widetilde{\mathbf{D}}\widetilde{\mathbf{R}} = \mathbf{D} = \begin{bmatrix} \mathbf{M} + \mathbf{A}\mathbf{N}^{-1}\mathbf{A}^T & \\ & \mathbf{N} + \mathbf{A}^T\mathbf{M}^{-1}\mathbf{A} \end{bmatrix}. \end{aligned}$$

Some properties of SQD matrices

By direct computation it is easy to prove that

$$\widetilde{\mathbf{C}}^2 = \begin{bmatrix} \mathbf{I}_m + \tilde{\mathbf{A}}\tilde{\mathbf{A}}^T & \\ & \mathbf{I}_n + \tilde{\mathbf{A}}^T\tilde{\mathbf{A}} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{D}}_1 & \\ & \widetilde{\mathbf{D}}_2 \end{bmatrix} = \widetilde{\mathbf{D}}.$$

$$\begin{aligned} \widetilde{\mathbf{C}}^{-1} &= \widetilde{\mathbf{D}}^{-1}\widetilde{\mathbf{C}} = \widetilde{\mathbf{C}}\widetilde{\mathbf{D}}^{-1}; \\ \widetilde{\mathbf{C}}\widetilde{\mathbf{D}} &= \widetilde{\mathbf{C}}^3 = \widetilde{\mathbf{D}}\widetilde{\mathbf{C}}; \\ \mathbf{C}\mathbf{H}^{-1}\mathbf{C} &= \widetilde{\mathbf{R}}^T\widetilde{\mathbf{D}}\widetilde{\mathbf{R}} = \mathbf{D} = \begin{bmatrix} \mathbf{M} + \mathbf{A}\mathbf{N}^{-1}\mathbf{A}^T & \\ & \mathbf{N} + \mathbf{A}^T\mathbf{M}^{-1}\mathbf{A} \end{bmatrix}. \end{aligned}$$

$$\left(\mathbf{H}^{-1}\mathbf{C}\right)^2 = \widetilde{\mathbf{R}}^{-1}\widetilde{\mathbf{C}}^2\widetilde{\mathbf{R}} = \widetilde{\mathbf{R}}^{-1}\widetilde{\mathbf{D}}\widetilde{\mathbf{R}} = \mathbf{H}^{-1}\mathbf{D},$$

$$\left(\mathbf{H}^{-1}\mathbf{C}\right)^3 = \widetilde{\mathbf{R}}^{-1}\widetilde{\mathbf{C}}^3\widetilde{\mathbf{R}} = \mathbf{H}^{-1}\mathbf{C}\mathbf{H}^{-1}\mathbf{D} = \mathbf{H}^{-1}\mathbf{D}\mathbf{H}^{-1}\mathbf{C}$$

$$\mathbf{C}^{-1} = \mathbf{D}^{-1}\mathbf{C}\mathbf{H}^{-1} = \mathbf{H}^{-1}\mathbf{C}\mathbf{D}^{-1}.$$

Some properties of SQD matrices

$\widetilde{\mathbf{D}}$ and $\widetilde{\mathbf{C}}$ commute.
Both matrices can be simultaneously diagonalized by the generalized eigenvalues of

$$\mathbf{C}\mathbf{z} = \lambda_j \mathbf{H}\mathbf{z},$$

where the $\lambda_j,\ j = 1, \ldots, p = \mathrm{rank}(\bar{\mathbf{A}})$ are the same eigenvalues of $\widehat{\mathbf{C}}$

Krylov subspaces

Hereafter we will denote by

$$\widetilde{\mathrm{K}}_i(\widetilde{\mathbf{C}}, \mathbf{z}) = \mathrm{Range}\Big\{\mathbf{z}, \widetilde{\mathbf{C}}\mathbf{z}, \widetilde{\mathbf{C}}^2\mathbf{z}, \ldots, \widetilde{\mathbf{C}}^{i-1}\mathbf{z}, \widetilde{\mathbf{C}}^i\mathbf{z}\Big\}$$

the Krylov subspace generated by $\widetilde{\mathbf{C}}$ and a vector $\mathbf{z}$. We point out that $\widetilde{\mathrm{K}}_i(\widetilde{\mathbf{C}}, \mathbf{z})$ are also the Krylov subspaces used to define the Lanczos algorithm applied to $\mathbf{C}$ symmetrically preconditioned by $\widetilde{\mathbf{R}}$.

$$\widetilde{\mathrm{K}}_i(\mathbf{H}^{-1}\mathbf{C}, \mathbf{w}) = \widetilde{\mathbf{R}}^{-1}\widetilde{\mathrm{K}}_i(\widetilde{\mathbf{C}}, \mathbf{z}), \;\; \text{where} \qquad \mathbf{w} = \widetilde{\mathbf{R}}\mathbf{z}.$$

Krylov subspaces

$$\left.\begin{array}{l} \widetilde{\mathbf{C}}^{2k} = \widetilde{\mathbf{D}}^{k} \\ \widetilde{\mathbf{C}}^{2k+1} = \widetilde{\mathbf{C}}\widetilde{\mathbf{D}}^{k} = \widetilde{\mathbf{D}}^{k}\widetilde{\mathbf{C}} \end{array}\right\}.$$

Therefore,

$$\begin{aligned} \widetilde{\mathsf{K}}_k(\widetilde{\mathbf{C}}, \mathbf{z}) &= \widetilde{\mathsf{K}}_{\lfloor k/2 \rfloor}(\widetilde{\mathbf{D}}, \mathbf{z}) + \widetilde{\mathsf{K}}_{\lceil k/2 \rceil - 1}(\widetilde{\mathbf{D}}, \widetilde{\mathbf{C}}\mathbf{z}) \\ &= \widetilde{\mathsf{K}}_{\lfloor k/2 \rfloor}(\widetilde{\mathbf{D}}, \mathbf{z}) + \widetilde{\mathbf{C}}\widetilde{\mathsf{K}}_{\lceil k/2 \rceil - 1}(\widetilde{\mathbf{D}}, \mathbf{z}). \end{aligned}$$

Krylov subspaces

Finally, denoting by $\widetilde{\mathbf{D}}_1$ and $\widetilde{\mathbf{D}}_2$ the diagonal blocks of $\widetilde{\mathbf{D}}$, i.e. we have:

$$\widetilde{\mathsf{K}}_i(\widetilde{\mathbf{D}}, \begin{bmatrix}\mathbf{z}^1\\ \mathbf{z}^2\end{bmatrix}) = \begin{bmatrix}\mathsf{K}_i(\widetilde{\mathbf{D}}_1, \mathbf{z}^1)\\ 0\end{bmatrix} \oplus \begin{bmatrix}0\\ \mathsf{K}_i(\widetilde{\mathbf{D}}_2, \mathbf{z}^2)\end{bmatrix}$$

and

$$\begin{aligned}\widetilde{\mathbf{C}}\widetilde{\mathsf{K}}_i(\widetilde{\mathbf{D}}, \begin{bmatrix}\mathbf{z}^1\\ \mathbf{z}^2\end{bmatrix}) &= \begin{bmatrix}\mathsf{K}_i(\widetilde{\mathbf{D}}_1, \mathbf{z}^1)\\ \tilde{\mathbf{A}}^T\mathsf{K}_i(\widetilde{\mathbf{D}}_1, \mathbf{z}^1)\end{bmatrix} \oplus \begin{bmatrix}\tilde{\mathbf{A}}\mathsf{K}_i(\widetilde{\mathbf{D}}_2, \mathbf{z}^2)\\ -\mathsf{K}_i(\widetilde{\mathbf{D}}_2, \mathbf{z}^2)\end{bmatrix}\\ &= \begin{bmatrix}\mathsf{K}_i(\widetilde{\mathbf{D}}_1, \mathbf{z}^1)\\ \mathsf{K}_i(\widetilde{\mathbf{D}}_2, \tilde{\mathbf{A}}^T\mathbf{z}^1)\end{bmatrix} \oplus \begin{bmatrix}\mathsf{K}_i(\widetilde{\mathbf{D}}_1, \tilde{\mathbf{A}}\mathbf{z}^2)\\ -\mathsf{K}_i(\widetilde{\mathbf{D}}_2, \mathbf{z}^2)\end{bmatrix}.\end{aligned}$$

Generalized Golub-Kahan bidiagonalization

TWO VARIANTS

Generalized Golub-Kahan bidiagonalization

$$\left\{ \begin{array}{lcl} \mathbf{A}\tilde{\mathbf{Q}} & = & \mathbf{M}\tilde{\mathbf{V}} \begin{bmatrix} \tilde{\mathbf{B}} \\ 0 \end{bmatrix} \\ \mathbf{A}^T\tilde{\mathbf{V}} & = & \mathbf{N}\tilde{\mathbf{Q}} \left[\tilde{\mathbf{B}}^T ; 0 \right] \end{array} \right. \qquad \begin{array}{l} \tilde{\mathbf{V}}^T\mathbf{M}\tilde{\mathbf{V}} = \mathbf{I}_m \\ \tilde{\mathbf{Q}}^T\mathbf{N}\tilde{\mathbf{Q}} = \mathbf{I}_n \end{array}$$

where

$$\tilde{\mathbf{B}} = \begin{bmatrix} \tilde{\alpha}_1 & 0 & 0 & \cdots & 0 \\ \tilde{\beta}_2 & \tilde{\alpha}_2 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & \tilde{\beta}_{n-1} & \tilde{\alpha}_{n-1} & 0 \\ 0 & \cdots & 0 & \tilde{\beta}_n & \tilde{\alpha}_n \\ 0 & \cdots & 0 & 0 & \tilde{\beta}_{n+1} \end{bmatrix}.$$

Generalized Golub-Kahan bidiagonalization

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

where

$$\mathbf{B} = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \beta_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & \alpha_{n-1} & \beta_{n-1} \\ 0 & \cdots & 0 & 0 & \alpha_n \end{bmatrix}.$$

Generalized Least Squares

Normal equations: $(\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}+\mathbf{N})\mathbf{y}=\mathbf{A}^T\mathbf{M}^{-1}\mathbf{b}$.

Generalized Least Squares

Normal equations: $(\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}+\mathbf{N})\mathbf{y}=\mathbf{A}^T\mathbf{M}^{-1}\mathbf{b}$.

At k-th iteration, seek $y \approx \mathbf{y}_k := \tilde{\mathbf{V}}_k\bar{\mathbf{y}}_k$:

$$(\tilde{\mathbf{B}}_k^T\tilde{\mathbf{B}}_k+\mathbf{I})\bar{\mathbf{y}}_k=\tilde{\mathbf{B}}_k^T\beta_1\mathbf{e}_1$$

Generalized Least Squares

Normal equations: $(\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}+\mathbf{N})\mathbf{y}=\mathbf{A}^T\mathbf{M}^{-1}\mathbf{b}$.

At k-th iteration, seek $y \approx \mathbf{y}_k := \tilde{\mathbf{V}}_k\bar{\mathbf{y}}_k$:

$$(\tilde{\mathbf{B}}_k^T\tilde{\mathbf{B}}_k+\mathbf{I})\bar{\mathbf{y}}_k=\tilde{\mathbf{B}}_k^T\beta_1\mathbf{e}_1$$

i.e.:

$$\min_{\bar{\mathbf{y}}\in\mathrm{R}^k} \ \frac{1}{2}\left\|\begin{bmatrix}\tilde{\mathbf{B}}_k\\ \mathbf{I}\end{bmatrix}\bar{\mathbf{y}}-\begin{bmatrix}\beta_1\mathbf{e}_1\\ 0\end{bmatrix}\right\|_2^2$$

Generalized Least Squares

Normal equations: $(\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}+\mathbf{N})\mathbf{y}=\mathbf{A}^T\mathbf{M}^{-1}\mathbf{b}$.

At k-th iteration, seek $y \approx \mathbf{y}_k := \tilde{\mathbf{V}}_k\bar{\mathbf{y}}_k$:

$$(\tilde{\mathbf{B}}_k^T\tilde{\mathbf{B}}_k+\mathbf{I})\bar{\mathbf{y}}_k=\tilde{\mathbf{B}}_k^T\beta_1\mathbf{e}_1$$

i.e.:

$$\min_{\bar{\mathbf{y}}\in\mathrm{R}^k} \ \frac{1}{2}\left\|\begin{bmatrix}\tilde{\mathbf{B}}_k\\ \mathbf{I}\end{bmatrix}\bar{\mathbf{y}}-\begin{bmatrix}\beta_1\mathbf{e}_1\\ 0\end{bmatrix}\right\|_2^2$$

or:

$$\begin{bmatrix}\mathbf{I} & \tilde{\mathbf{B}}_k\\ \tilde{\mathbf{B}}_k^T & -\mathbf{I}\end{bmatrix}\begin{bmatrix}\bar{\mathbf{x}}_k\\ \bar{\mathbf{y}}_k\end{bmatrix}=\begin{bmatrix}\beta_1\mathbf{e}_1\\ 0\end{bmatrix}.$$

Generalized LSQR

Solve

$$\min_{\bar{\mathbf{y}} \in \mathrm{R}^k} \ \frac{1}{2} \left\| \begin{bmatrix} \tilde{\mathbf{B}}_k \\ \mathbf{I} \end{bmatrix} \bar{\mathbf{y}} - \begin{bmatrix} \beta_1 \mathbf{e}_1 \\ 0 \end{bmatrix} \right\|_2^2$$

by specialized Givens Rotations (Eliminate $\mathbf{I}$ first and $\tilde{\mathbf{R}}_k$ will be upper bidiagonal)

Generalized LSQR

Solve

$$\min_{\bar{\mathbf{y}} \in \mathrm{R}^k} \ \frac{1}{2} \left\| \begin{bmatrix} \tilde{\mathbf{B}}_k \\ \mathbf{I} \end{bmatrix} \bar{\mathbf{y}} - \begin{bmatrix} \beta_1 \mathbf{e}_1 \\ 0 \end{bmatrix} \right\|_2^2$$

by specialized Givens Rotations (Eliminate $\mathbf{I}$ first and $\tilde{\mathbf{R}}_k$ will be upper bidiagonal)

$$\min_{\bar{\mathbf{y}} \in \mathrm{R}^k} \ \frac{1}{2} \left\| \begin{bmatrix} \tilde{\mathbf{R}}_k \\ 0 \end{bmatrix} \bar{\mathbf{y}} - \begin{bmatrix} \phi_k \\ 0 \end{bmatrix} \right\|_2^2 .$$

Generalized LSQR

Solve

$$\min_{\bar{\mathbf{y}}\in \mathrm{R}^k} \ \frac{1}{2}\left\| \begin{bmatrix} \tilde{\mathbf{B}}_k \\ \mathbf{I} \end{bmatrix} \bar{\mathbf{y}} - \begin{bmatrix} \beta_1 \mathbf{e}_1 \\ 0 \end{bmatrix} \right\|_2^2$$

by specialized Givens Rotations (Eliminate $\mathbf{I}$ first and $\tilde{\mathbf{R}}_k$ will be upper bidiagonal)

$$\min_{\bar{\mathbf{y}}\in \mathrm{R}^k} \ \frac{1}{2}\left\| \begin{bmatrix} \tilde{\mathbf{R}}_k \\ 0 \end{bmatrix} \bar{\mathbf{y}} - \begin{bmatrix} \phi_k \\ 0 \end{bmatrix} \right\|_2^2 .$$

As in Paige-Saunders '82 we can build recursive expressions of $\mathbf{y}_k$

$$\mathbf{y}_{k+1} = \mathbf{y}_k + \mathbf{d}_k \phi_k \quad \left(\mathbf{D}_k = \tilde{\mathbf{V}}_k \tilde{\mathbf{R}}_k^{-1} \right)$$

and we have that

$$||\bar{\mathbf{y}}||^2_{\mathbf{N}+\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}} = \sum_{j=1}^{m} \phi_j^2 \quad \text{and} \quad ||\bar{\mathbf{y}} - \mathbf{y}_k||^2_{\mathbf{N}+\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}} = \sum_{j=k+1}^{m} \phi_j^2$$

Error bound

Lower bound We can estimate $||\bar{\mathbf{y}} - \mathbf{y}_k||^2_{\mathbf{N}+\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}}$ by the lower bound

$$\xi^2_{k,d} = \sum_{j=k+1}^{k+d+1} \phi_j^2 < ||\bar{\mathbf{y}} - \mathbf{y}_k||^2_{\mathbf{N}+\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}}.$$

and $||\bar{\mathbf{y}}||^2_{\mathbf{N}+\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}}$ by the lower bound $\sum_{j=1}^{k} \phi_j^2$. Given a threshold $\tau < 1$ and an integer d, we can stop the iterations when

$$\xi^2_{k,d} \leq \tau \sum_{j=1}^{k+d+1} \phi_j^2 < \tau \sum_{j=1}^{k} \phi_j^2 < \tau ||\bar{\mathbf{y}}||^2_{\mathbf{N}+\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}}.$$

Error bound

Lower bound We can estimate $||\bar{\mathbf{y}} - \mathbf{y}_k||^2_{\mathbf{N}+\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}}$ by the lower bound

$$\xi^2_{k,d} = \sum_{j=k+1}^{k+d+1} \phi_j^2 < ||\bar{\mathbf{y}} - \mathbf{y}_k||^2_{\mathbf{N}+\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}}.$$

and $||\bar{\mathbf{y}}||^2_{\mathbf{N}+\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}}$ by the lower bound $\sum_{j=1}^{k} \phi_j^2$. Given a threshold $\tau < 1$ and an integer d, we can stop the iterations when

$$\xi^2_{k,d} \leq \tau \sum_{j=1}^{k+d+1} \phi_j^2 < \tau \sum_{j=1}^{k} \phi_j^2 < \tau ||\bar{\mathbf{y}}||^2_{\mathbf{N}+\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}}.$$

Upper bound Despite being very inexpensive, the previous estimator is still a lower bound of the error. We can use an approach inspired by the Gauss-Radau quadrature algorithm and similar to the one described in Golub-Meurant (2010).

Generalized CRAIG

$$\min_{\mathbf{y},\mathbf{x}} \ \tfrac{1}{2}(\|\mathbf{y}\|^2_{\mathbf{N}} + \|\mathbf{x}\|^2_{\mathbf{M}}) \quad \text{s.t. } \mathbf{A}\mathbf{y} + \mathbf{M}\mathbf{x} = \mathbf{b}.$$

Generalized CRAIG

$$\min_{\mathbf{y},\mathbf{x}} \ \tfrac{1}{2}(\|\mathbf{y}\|^2_{\mathbf{N}} + \|\mathbf{x}\|^2_{\mathbf{M}}) \quad \text{s.t. } \mathbf{A}\mathbf{y} + \mathbf{M}\mathbf{x} = \mathbf{b}.$$

At step k of GK bidiagonalization, we seek

$$\mathbf{x} \approx \mathbf{x}_k := \mathbf{U}_k\bar{\mathbf{x}}_k, \qquad \text{and} \qquad \mathbf{y} \approx \mathbf{y}_k := \mathbf{V}_k\bar{\mathbf{y}}_k.$$

Generalized CRAIG

$$\min_{\mathbf{y},\mathbf{x}} \ \tfrac{1}{2}(\|\mathbf{y}\|_{\mathbf{N}}^2 + \|\mathbf{x}\|_{\mathbf{M}}^2) \quad \text{s.t. } \mathbf{A}\mathbf{y} + \mathbf{M}\mathbf{x} = \mathbf{b}.$$

At step k of GK bidiagonalization, we seek

$$\mathbf{x} \approx \mathbf{x}_k := \mathbf{U}_k\bar{\mathbf{x}}_k, \qquad \text{and} \qquad \mathbf{y} \approx \mathbf{y}_k := \mathbf{V}_k\bar{\mathbf{y}}_k.$$

$$\min_{\bar{\mathbf{y}},\bar{\mathbf{x}}} \ \tfrac{1}{2}(\|\bar{\mathbf{y}}\|^2 + \|\bar{\mathbf{x}}\|^2) \quad \text{s.t. } \mathbf{B}_k\bar{\mathbf{y}}_k + \bar{\mathbf{x}}_k = \beta_1\mathbf{e}_1$$

Generalized CRAIG

$$\min_{\mathbf{y},\mathbf{x}} \ \tfrac{1}{2}(\|\mathbf{y}\|^2_{\mathbf{N}} + \|\mathbf{x}\|^2_{\mathbf{M}}) \quad \text{s.t. } \mathbf{A}\mathbf{y} + \mathbf{M}\mathbf{x} = \mathbf{b}.$$

At step k of GK bidiagonalization, we seek

$$\mathbf{x} \approx \mathbf{x}_k := \mathbf{U}_k\bar{\mathbf{x}}_k, \qquad \text{and} \qquad \mathbf{y} \approx \mathbf{y}_k := \mathbf{V}_k\bar{\mathbf{y}}_k.$$

$$\min_{\bar{\mathbf{y}},\bar{\mathbf{x}}} \ \tfrac{1}{2}(\|\bar{\mathbf{y}}\|^2 + \|\bar{\mathbf{x}}\|^2) \quad \text{s.t. } \mathbf{B}_k\bar{\mathbf{y}}_k + \bar{\mathbf{x}}_k = \beta_1\mathbf{e}_1$$

or:

$$\min_{\bar{\mathbf{y}}\in\mathrm{R}^k} \ \tfrac{1}{2}\left\| \begin{bmatrix} \mathbf{B}_k \\ \mathbf{I} \end{bmatrix} \bar{\mathbf{y}} - \begin{bmatrix} \beta_1\mathbf{e}_1 \\ 0 \end{bmatrix} \right\|_2^2 .$$

Generalized CRAIG

By contrast with generalized LSQR, we solve the SQD subsystem

$$\begin{bmatrix} \mathbf{I}_k & \mathbf{B}_k \\ \mathbf{B}_k^T & -I_k \end{bmatrix} \begin{bmatrix} \bar{\mathbf{x}}_k \\ \bar{\mathbf{y}}_k \end{bmatrix} = \begin{bmatrix} \beta_1 \mathbf{e}_1 \\ 0 \end{bmatrix}$$

Generalized CRAIG

By contrast with generalized LSQR, we solve the SQD subsystem

$$\begin{bmatrix} \mathbf{I}_k & \mathbf{B}_k \\ \mathbf{B}_k^T & -I_k \end{bmatrix} \begin{bmatrix} \bar{\mathbf{x}}_k \\ \bar{\mathbf{y}}_k \end{bmatrix} = \begin{bmatrix} \beta_1 \mathbf{e}_1 \\ 0 \end{bmatrix}$$

Following Saunders (1995) and Paige (1974), we compute an LQ factorization to the k-by-$2k$ matrix $\begin{bmatrix} \mathbf{B}_k & \mathbf{I}_k \end{bmatrix}$ by applying $2k - 1$ Givens rotations that zero out the identity block.

Generalized CRAIG

By contrast with generalized LSQR, we solve the SQD subsystem

$$\begin{bmatrix} \mathbf{I}_k & \mathbf{B}_k \\ \mathbf{B}_k^T & -I_k \end{bmatrix} \begin{bmatrix} \bar{\mathbf{x}}_k \\ \bar{\mathbf{y}}_k \end{bmatrix} = \begin{bmatrix} \beta_1 \mathbf{e}_1 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{B}_k & \mathbf{I}_k \end{bmatrix} \mathbf{Q}_k^T = \begin{bmatrix} \hat{\mathbf{B}}_k & 0 \end{bmatrix} \qquad \mathbf{Q}_k^T \mathbf{Q}_k = \mathbf{I}$$

where

$$\hat{\mathbf{B}}_k := \begin{bmatrix} \hat{\alpha}_1 & & & \\ \hat{\beta}_2 & \hat{\alpha}_2 & & \\ & \ddots & \ddots & \\ & & \hat{\beta}_k & \hat{\alpha}_k \end{bmatrix}.$$

Generalized CRAIG

$$\beta_1 \mathbf{e}_1 = \mathbf{B}_k \bar{\mathbf{y}}_k + \bar{\mathbf{x}}_k = \begin{bmatrix} \mathbf{B}_k & I_k \end{bmatrix} \begin{bmatrix} \bar{\mathbf{y}}_k \\ \bar{\mathbf{x}}_k \end{bmatrix} =$$

$$\begin{bmatrix} \hat{\mathbf{B}}_k & 0 \end{bmatrix} \mathbf{Q}_k \begin{bmatrix} \bar{\mathbf{y}}_k \\ \bar{\mathbf{x}}_k \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{B}}_k & 0 \end{bmatrix} \begin{bmatrix} \bar{\mathbf{z}}_k \\ 0 \end{bmatrix} = \hat{\mathbf{B}}_k \bar{\mathbf{z}}_k,$$

for some $\bar{\mathbf{z}}_k \in \mathrm{R}^k$: $\bar{\mathbf{z}}_k = (\zeta_1, \ldots, \zeta_k)$

Generalized CRAIG

$$\beta_1 \mathbf{e}_1 = \mathbf{B}_k \bar{\mathbf{y}}_k + \bar{\mathbf{x}}_k = \begin{bmatrix} \mathbf{B}_k & I_k \end{bmatrix} \begin{bmatrix} \bar{\mathbf{y}}_k \\ \bar{\mathbf{x}}_k \end{bmatrix} =$$

$$\begin{bmatrix} \hat{\mathbf{B}}_k & 0 \end{bmatrix} \mathbf{Q}_k \begin{bmatrix} \bar{\mathbf{y}}_k \\ \bar{\mathbf{x}}_k \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{B}}_k & 0 \end{bmatrix} \begin{bmatrix} \bar{\mathbf{z}}_k \\ 0 \end{bmatrix} = \hat{\mathbf{B}}_k \bar{\mathbf{z}}_k,$$

for some $\bar{\mathbf{z}}_k \in \mathrm{R}^k$: $\bar{\mathbf{z}}_k = (\zeta_1, \ldots, \zeta_k)$

$$\zeta_1 = \beta_1/\hat{\alpha}_1, \quad \zeta_{i+1} = -\hat{\beta}_{i+1}\zeta_i/\hat{\alpha}_{i+1}, \quad (i = 1, \ldots, k-1).$$

Generalized CRAIG

Solving for $\mathbf{x}_k$ directly, and bypassing $\bar{\mathbf{x}}_k$, may now be done. By definition,

$$\mathbf{x}_k = \mathbf{U}_k \bar{\mathbf{x}}_k = \mathbf{U}_k \hat{\mathbf{B}}_k^{-T} \bar{\mathbf{z}}_k.$$

Since $\hat{\mathbf{B}}_k^{-T}$ is upper bidiagonal, all components of $\hat{\mathbf{B}}_k^{-T} \bar{\mathbf{z}}_k$ are likely to change at every iteration. Fortunately, upon defining $\mathbf{D}_k := \mathbf{U}_k \hat{\mathbf{B}}_k^{-T}$, and denoting $\mathbf{d}_i$ the i-th column of $\mathbf{D}_k$, we are able to use a recursion formula for $\mathbf{x}_k$ provided that $\mathbf{d}_i$ may be found easily. Slightly rearranging, we have

$$\hat{\mathbf{B}}_k \mathbf{D}_k^{\mathsf{T}} = \mathbf{U}_k^{\mathsf{T}}$$

and therefore it is easy to identify each $\mathbf{d}_i$—i.e., each row of $\mathbf{D}_k^{\mathsf{T}}$—recursively.

Generalized CRAIG

Solving for $\mathbf{x}_k$ directly, and bypassing $\bar{\mathbf{x}}_k$, may now be done. By definition,

$$\mathbf{x}_k = \mathbf{U}_k \bar{\mathbf{x}}_k = \mathbf{U}_k \hat{\mathbf{B}}_k^{-T} \bar{\mathbf{z}}_k.$$

$$\mathbf{d}_1 := \mathbf{u}_1/\hat{\alpha}_1, \quad \mathbf{d}_{i+1} := (\mathbf{u}_{i+1} - \hat{\beta}_{i+1}\mathbf{d}_i)/\hat{\alpha}_{i+1}, \quad (i = 1, \ldots, k-1).$$

This yields the update

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \zeta_{k+1}\mathbf{d}_{k+1}$$

for $\mathbf{x}_{k+1}$.

Generalized CRAIG: errors bound

Let $\hat{\mathbf{B}}_k$ be defined as above and $\mathbf{D}_k := \mathbf{U}_k \hat{\mathbf{B}}_k^{-T}$. For $k = 1, \ldots, n$, we have

$$\mathbf{D}_k^{\mathsf{T}} (\mathbf{A}\mathbf{N}^{-1}\mathbf{A}^{\mathsf{T}} + \mathbf{M}) \mathbf{D}_k = \mathbf{I}_k.$$

In particular,

$$\mathbf{x}_k = \sum_{j=1}^{k} \zeta_j \mathbf{d}_j$$

and we have the estimates

$$\|\mathbf{x}_k\|^2_{\mathbf{A}\mathbf{N}^{-1}\mathbf{A}^{\mathsf{T}} + \mathbf{M}} = \sum_{i=1}^{k} \zeta_i^2, \tag{4a}$$

$$\|\mathbf{x}^* - \mathbf{x}_k\|^2_{\mathbf{A}\mathbf{N}^{-1}\mathbf{A}^{T} + \mathbf{M}} = \sum_{i=k+1}^{n} \zeta_i^2, \tag{4b}$$

Generalized CRAIG: errors bound

As for generalized LSQR, we can estimate the error using the windowing technique and we can give a lower bound of the error by

$$\xi_{k,d}^2 = \sum_{j=k+1}^{k+d+1} \zeta_i^2 \leq \|\mathbf{x}^* - \mathbf{x}_k\|^2_{\mathbf{A}\mathbf{N}^{-1}\mathbf{A}^T+\mathbf{M}}$$

and we can estimate $\|\mathbf{x}^*\|_{\mathbf{A}\mathbf{N}^{-1}\mathbf{A}^T+\mathbf{M}}$ by the lower bound $\sum_{j=1}^{k} \zeta_j^2$.

Generalized CRAIG: errors bound

As for GLSQR. If we know a lower bound of singular values we can use an approach inspired by the Gauss-Radau quadrature algorithm and similar to the one described in Golub-Meurant (2010).

Other variants:

Generalized LSMR

$$\underset{\mathbf{y}\in\mathbb{R}^m}{\text{minimize}} \;\; \frac{1}{2}\|\mathbf{N}^{-\frac{1}{2}}(\mathbf{A}^T\mathbf{M}^{-1}\mathbf{b}-(\mathbf{A}^T\mathbf{M}^{-1}\mathbf{A}+\mathbf{N})\mathbf{y}))\|_2.$$

Error bounds similar to the ones given above exist for the MR variants

Other variants:

Generalized LSMR

Generalized Craig-MR

Error bounds similar to the ones given above exist for the MR variants

Numerical experiments

We will focus on optimization problems:

$$\underset{\mathbf{x}\in\mathrm{R}^n}{\text{minimize}}\ \mathbf{g}^T\mathbf{x} + \tfrac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} \quad \text{subject to}\ \mathbf{C}\mathbf{x} = \mathbf{d},\ \mathbf{x} \geq 0,$$

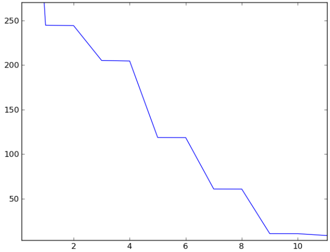
where $\mathbf{g} \in \mathrm{R}^n$ and $\mathbf{H} = \mathbf{H}^T \in \mathrm{R}^{n\times n}$ is positive semi-definite, and result in linear systems with coefficient matrix

$$\begin{bmatrix} \mathbf{H} + \mathbf{X}^{-1}\mathbf{Z} + \rho\mathbf{I} & \mathbf{C}^T \\ \mathbf{C} & -\delta\mathbf{I} \end{bmatrix}$$

where $\rho > 0$ and $\delta > 0$ are regularization parameters.

Numerical experiments MINRES

This is a blow-up of some iterations



Numerical experiments GLSQR

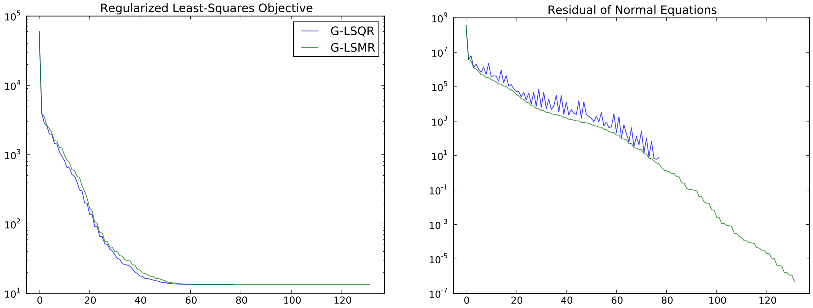


Figure : Problem DUAL1 $(255, 171)$.

Numerical experiments GLSQR

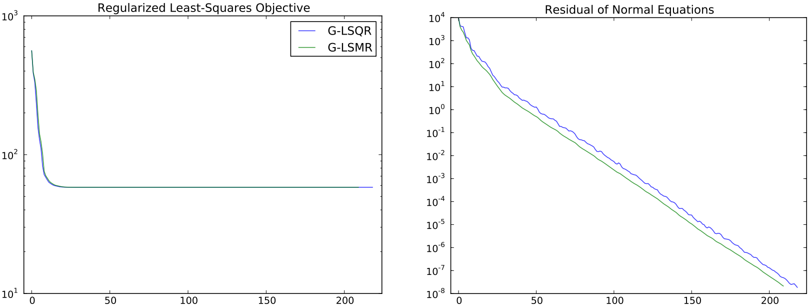


Figure : Problem MOSARQP1 $(5700, 3200)$.

How to choose d?

problem	m	n
dual1	255	171
stcqp1	12291	10246
qpcboei1	1355	980

Numerical experiments GCraig

$d = 5, 15$

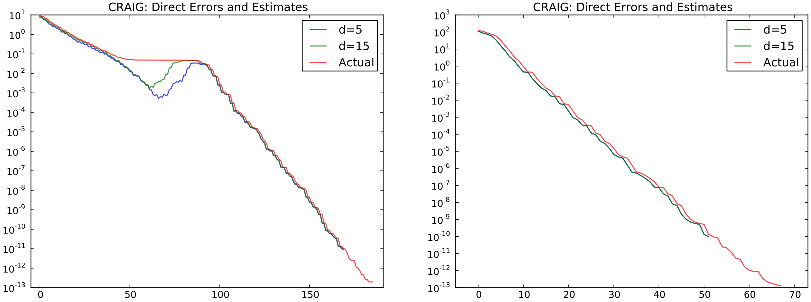


Figure : Problem dual1

CG?

Numerical experiments CG

$d = 5, 15$

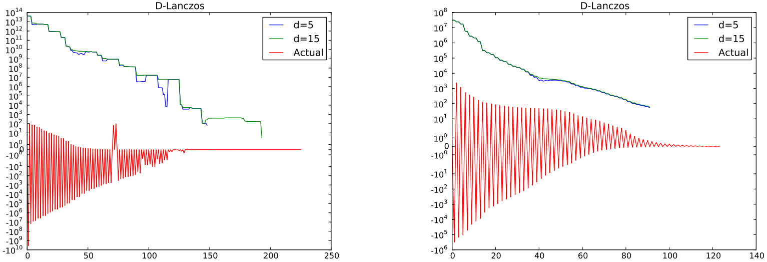


Figure : Problem DUAL1 and MOSARQP1 $(5700, 3200)$.

Numerical experiments CG

$d = 5, 15$

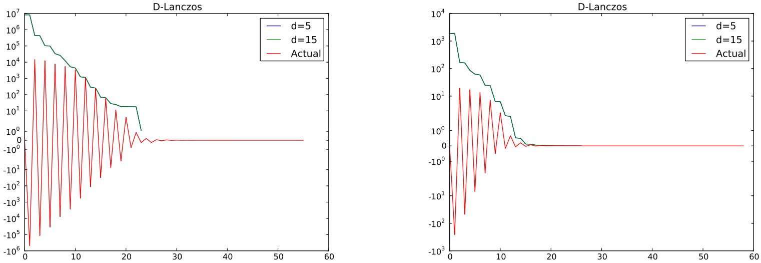


Figure : Problem Stokes (IFISS 3.1): colliding and cavity

Conclusions

- Preconditioning $\longrightarrow$ Norms i.e. different topologies!!

Conclusions

- Preconditioning $\longrightarrow$ Norms i.e. different topologies!!
- Nice relation between the algebraic error and the approximation error

Conclusions

- Preconditioning $\longrightarrow$ Norms i.e. different topologies!!
- Nice relation between the algebraic error and the approximation error
- A. and Orban "Iterative methods for symmetric quasi definite systems" Cahier du GERAD G-2013-32

Lecture on linear regression and LS

- QR algorithm
- Sparse least-squares problems
- Rounding error analysis

Elementary matrices

- Givens transformation:

$$\mathbf{G} = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & c & & s & & \\ & & & \ddots & & & \\ & & -s & & c & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix}$$

Elementary matrices

- Givens transformation:

$$\mathbf{G} = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & c & & s & & \\ & & & \ddots & & & \\ & & -s & & c & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \qquad c^2 + s^2 = 1$$

Elementary matrices

- Givens transformation:

$$\mathbf{G} = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & c & & s & & \\ & & & \ddots & & & \\ & & -s & & c & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \qquad c^2 + s^2 = 1$$

- Householder transformation

$$\mathbf{H} = \mathbf{I} - 2\frac{\mathbf{y}\mathbf{y}^T}{\mathbf{y}^T\mathbf{y}}$$

Givens transformation

$\mathbf{u} \in \mathrm{R}^n$ find $n-1$, $\mathbf{G}_i$ such that

$$\mathbf{G}_{n-1} \dots \mathbf{G}_1 \mathbf{u} = ||\mathbf{u}||_2 \mathbf{e}_1$$

Givens transformation

$\mathbf{u} \in \mathrm{R}^n$ find $n-1$, $\mathbf{G}_i$ such that

$$\mathbf{G}_{n-1} \dots \mathbf{G}_1 \mathbf{u} = ||\mathbf{u}||_2 \mathbf{e}_1$$

If $n = 2$

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \sqrt{x^2 + y^2} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Givens transformation

$\mathbf{u} \in \mathrm{R}^n$ find $n-1$, $\mathbf{G}_i$ such that

$$\mathbf{G}_{n-1} \dots \mathbf{G}_1 \mathbf{u} = ||\mathbf{u}||_2 \mathbf{e}_1$$

If $n = 2$

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \sqrt{x^2 + y^2} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$c = \frac{x}{\sqrt{x^2 + y^2}} \qquad s = \frac{y}{\sqrt{x^2 + y^2}}$$

Householder transformation

$$(\mathbf{I} - 2\frac{\mathbf{y}\mathbf{y}^T}{\mathbf{y}^T\mathbf{y}})\mathbf{u} = \pm||\mathbf{u}||_2\mathbf{e}_1$$

Householder transformation

$$(\mathbf{I} - 2\frac{\mathbf{y}\mathbf{y}^T}{\mathbf{y}^T\mathbf{y}})\mathbf{u} = \pm||\mathbf{u}||_2\mathbf{e}_1$$

$$\mathbf{y} = \mathbf{u} \pm ||\mathbf{u}||_2\mathbf{e}_1$$

Householder transformation

$$(\mathbf{I} - 2\frac{\mathbf{y}\mathbf{y}^T}{\mathbf{y}^T\mathbf{y}})\mathbf{u} = \pm||\mathbf{u}||_2\mathbf{e}_1$$

$$\mathbf{y} = \mathbf{u} \pm ||\mathbf{u}||_2\mathbf{e}_1$$

$$\mathbf{y} = \mathbf{u} + sign(u_1)||\mathbf{u}||_2\mathbf{e}_1$$

to avoid cancellation

Householder transformation:example

$$\begin{aligned}
\mathbf{H} &= \mathbf{H}_m \ldots \mathbf{H}_2\mathbf{H}_1\mathbf{A} \\
\mathbf{HA} &= \begin{bmatrix} \overline{\mathbf{R}} \\ 0 \end{bmatrix} \\
\mathbf{H}^T\mathbf{H} &= \mathbf{H}\mathbf{H}^T = \mathbf{I}
\end{aligned}$$

Product of Householder transformations

Let $\mathbf{H}_1$ and $\mathbf{H}_2$ be two Householder matrices

$$\begin{aligned} \mathbf{H}_1 &= \mathbf{I} - \mathbf{y}\mathbf{y}^T \qquad \mathbf{H}_2 = \mathbf{I} - \mathbf{w}\mathbf{w}^T \\ \|\mathbf{y}\|_2 &= \sqrt{2} \qquad \|\mathbf{w}\|_2 = \sqrt{2} \end{aligned}$$

$$\begin{aligned} \mathbf{H}_1\mathbf{H}_2 &= (\mathbf{I} - \mathbf{y}\mathbf{y}^T)(\mathbf{I} - \mathbf{w}\mathbf{w}^T) \\ &= \mathbf{I} - \mathbf{y}\mathbf{y}^T - \mathbf{w}\mathbf{w}^T + \mathbf{y}\mathbf{y}^T\mathbf{w}\mathbf{w}^T \\ &= \mathbf{I} - \left[\begin{array}{cc} \mathbf{y} & \mathbf{w} \end{array}\right] \left[\begin{array}{cc} 1 & -\mathbf{y}^T\mathbf{w} \\ 0 & 1 \end{array}\right] \left[\begin{array}{c} \mathbf{y}^T \\ \mathbf{w}^T \end{array}\right] \\ &= \mathbf{I} - \mathbf{Y}\mathbf{T}\mathbf{Y}^T \end{aligned}$$

Product of Householder transformations

Let $\mathbf{H}_1$ and $\mathbf{H}_2$ be two products of Householder matrices

$$\mathbf{H}_1 = \mathbf{I} - \mathbf{Y}\mathbf{T}_1\mathbf{Y}^T \qquad \mathbf{H}_2 = \mathbf{I} - \mathbf{W}\mathbf{T}_2\mathbf{W}^T$$

$$\begin{aligned}
\mathbf{H}_1\mathbf{H}_2 &= (\mathbf{I} - \mathbf{Y}\mathbf{T}_1\mathbf{Y}^T)(\mathbf{H}_2 = \mathbf{I} - \mathbf{W}\mathbf{T}_2\mathbf{W}^T) \\
&= \mathbf{I} - \mathbf{Y}\mathbf{T}_1\mathbf{Y}^T - \mathbf{W}\mathbf{T}_2\mathbf{W}^T + \mathbf{Y}\mathbf{T}_1\mathbf{Y}^T\mathbf{W}\mathbf{T}_2\mathbf{W}^T \\
&= \mathbf{I} - \left[\begin{array}{cc} \mathbf{Y} & \mathbf{W} \end{array}\right] \left[\begin{array}{cc} \mathbf{T}_1 & -\mathbf{T}_1\mathbf{Y}^T\mathbf{W}\mathbf{T}_2 \\ 0 & \mathbf{T}_2 \end{array}\right] \left[\begin{array}{c} \mathbf{Y}^T \\ \mathbf{W}^T \end{array}\right] \\
&= \mathbf{I} - \mathbf{Y}_3\mathbf{T}_3\mathbf{Y}_3^T
\end{aligned}$$

BLAS-3 Operations in applying $\mathbf{I} - \mathbf{Y}_3\mathbf{T}_3\mathbf{Y}_3^T$ to a matrix.

Least Squares

$$\min_x ||\mathbf{Ax} - \mathbf{b}||_2$$

$|| \bullet ||_2$ invariant for orthonormal transformation

$$\begin{aligned}
\min_x ||\mathbf{Ax} - \mathbf{b}||_2 &= \min_x ||\mathbf{H}(\mathbf{Ax} - \mathbf{b})||_2 \\
&= \min_x || \begin{bmatrix} \overline{\mathbf{R}} \\ 0 \end{bmatrix} \mathbf{x} - \begin{bmatrix} \tilde{\mathbf{b}}_1 \\ \tilde{\mathbf{b}}_2 \end{bmatrix} ||_2 \\
&= \min_x || \begin{bmatrix} \overline{\mathbf{R}}\mathbf{x} - \tilde{\mathbf{b}}_1 \\ -\tilde{\mathbf{b}}_2 \end{bmatrix} ||_2
\end{aligned}$$

$$\mathbf{x} = \overline{\mathbf{R}}^{-1}\tilde{\mathbf{b}}_1$$

Error Analysis in mixed precision arithmetic

$$||fl(\mathbf{H}_i) - \mathbf{H}_i||_F \leq \epsilon + \mathcal{O}(\epsilon^2)$$

Let $\mathbf{C} \in \mathrm{R}^{m\times n}$. We first compute $\mathbf{B} = \mathbf{H}_i\mathbf{C}$ and let $\tilde{\mathbf{B}} = fl(\mathbf{H}_i\mathbf{C})$

$$\begin{aligned} ||\tilde{\mathbf{B}} - \mathbf{B}||_F &= ||[fl(fl(\mathbf{H}_i)\mathbf{C}) - fl(\mathbf{H}_i)\mathbf{C}] + (fl(\mathbf{H}_i)\mathbf{C} - \mathbf{H}_i\mathbf{C})||_F \\ &\leq ||fl(fl(\mathbf{H}_i)\mathbf{C}) - fl(\mathbf{H}_i)\mathbf{C}||_F + ||(fl(\mathbf{H}_i)\mathbf{C} - \mathbf{H}_i\mathbf{C})||_F \\ &= ||\mathbf{E}||_F + ||fl(\mathbf{H}_i) - \mathbf{H}_i||_F||\mathbf{C}||_F \end{aligned}$$

$$||\mathbf{E}||_F \leq \epsilon||\mathbf{C}||_F + \mathcal{O}(\epsilon^2)$$

$$||\tilde{\mathbf{B}} - \mathbf{B}||_F \leq c_1\epsilon||\mathbf{C}||_F + \mathcal{O}(\epsilon^2)$$

Error Analysis in mixed precision arithmetic

$$\mathbf{A}_1 = \mathbf{A} \qquad \mathbf{A}_{i+1} = \hat{\mathbf{H}}_i \mathbf{A}_i \qquad i = 1, \ldots, m$$

$\hat{\mathbf{H}}_i$ produces zeros in positions $i+1$ through m of column i of $\hat{\mathbf{H}}_i \mathbf{A}_i$
The computed quantities will be

$$\tilde{\mathbf{A}}_1 = \mathbf{A} \qquad \tilde{\mathbf{A}}_{i+1} = fl(\tilde{\mathbf{H}}_i \tilde{\mathbf{A}}_i) \qquad i = 1, \ldots, m$$

$\tilde{\mathbf{H}}_i = fl(\mathbf{H}_i)$ where $\mathbf{H}_i$ (orthonormal) would have produced zeros in positions $i+1$ through m of column i of $\mathbf{H}_i \tilde{\mathbf{A}}_i$

$$||\mathbf{H}_i \tilde{\mathbf{A}}_i - fl(\tilde{\mathbf{H}}_i \tilde{\mathbf{A}}_i)||_F \leq c_2 \epsilon ||\tilde{\mathbf{A}}_{i+1}||_F + \mathcal{O}(\epsilon^2)$$

$$||\tilde{\mathbf{A}}_m - \mathbf{H}_m \ldots \mathbf{H}_1 \mathbf{A}||_F \leq c_1 m \epsilon ||\mathbf{A}||_F + \mathcal{O}(\epsilon^2)$$

$$\tilde{\mathbf{A}}_m = \begin{bmatrix} \tilde{\mathbf{R}} \\ 0 \end{bmatrix}$$

Error Analysis in mixed precision arithmetic

Exists an orthonormal matrix $\mathbf{Q} = \mathbf{H}_m \dots \mathbf{H}_1$ and a matrix $\mathbf{E}$ such that

$$\begin{aligned} \mathbf{A} + \mathbf{E} &= \mathbf{Q} \begin{bmatrix} \tilde{\mathbf{R}} \\ 0 \end{bmatrix} \\ ||\mathbf{E}||_F &\leq c_2 m ||\mathbf{A}||_F \epsilon + \mathcal{O}(\epsilon^2) \end{aligned}$$

The computed solution $\tilde{\mathbf{x}}$ of the least-squares problem is the exact solution of the problem

$$\begin{aligned} \min_x \quad & ||(\mathbf{A} + \mathbf{E}_1)\mathbf{x} - (\mathbf{b} + \mathbf{g})||_2 = ||(\mathbf{A} + \mathbf{E}_1)\tilde{\mathbf{x}} - (\mathbf{b} + \mathbf{g})||_2 \\ ||\mathbf{E}_1||_F \quad \leq \quad & c_3 m ||\mathbf{A}||_F \epsilon + \mathcal{O}(\epsilon^2) \\ ||\mathbf{g}||_2 \quad \leq \quad & c_4 m ||\mathbf{b}||_2 \epsilon + \mathcal{O}(\epsilon^2) \end{aligned}$$

Linear regression

For any random vector $\mathbb{z}$, we denote by $E[\mathbb{z}]$ its mean and by $V[\mathbb{z}] = E[(\mathbb{z} - E[\mathbb{z}])(\mathbb{z} - E[\mathbb{z}])^T]$ its covariance matrix.
The notation $\mathbb{z} \sim \mathcal{N}\big(z, \mathbf{C}\big)$ means that $\mathbb{z}$ follows a Gaussian distribution with mean z and covariance matrix $\mathbf{C}$.

Linear regression

Let $\mathbf{A} \in \mathrm{R}^{m\times n}, \ \ m \geq n$, with $\mathrm{Rank}(\mathbf{A}) = n$. We consider the linear regression model

$$\mathbb{y} = \mathbf{A}\mathbb{x} + \mathbb{e},$$

where $E[\mathbb{e}] = 0$ and $V[\mathbb{e}] = \sigma^2 I_m$. We point out that $\mathbf{A}$ defines a given model and $\mathbb{x}$ is an unknown deterministic value.

Linear regression: Gauss-Markov Theorem

The minimum-variance unbiased (MVU) estimator of $\mathfrak{x}$ is related to $\mathbb{y}$ by the Gauss-Markov theorem.

Linear regression: Gauss-Markov Theorem

For the linear model the minimum-variance unbiased estimator of $\mathfrak{x}$ is given by

$$\mathbb{x}^* = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbb{y}.$$

$V[\mathbb{x}^*]$ satisfies $V[\mathbb{x}^*] = \sigma^2(\mathbf{A}^T\mathbf{A})^{-1}$. If in addition, $\mathbb{e} \sim \mathcal{N}\big(0, \sigma^2\mathbf{I}_m\big)$, $m > n$, and if we set

$$\mathbb{s}^2 = \frac{1}{m-n}||\mathbb{r}||_2^2,$$

where $\mathbb{r} = \mathbb{y} - \mathbf{A}\mathbb{x}^*$, we have for our estimator of $\mathfrak{x}$

$$\mathbb{x}^* \sim \mathcal{N}\big(\mathfrak{x}, \sigma^2(\mathbf{A}^T\mathbf{A})^{-1}\big),$$

and for $\mathbb{s}^2$, our estimator for σ^2,

$$\mathbb{s}^2 \sim \frac{\sigma^2}{m-n}\chi^2(m-n).$$

Linear regression: Gauss-Markov Theorem

Moreover, the predicted value $\hat{\mathbb{y}} = \mathbf{A}\mathbb{x}^*$ and the residual $\mathbb{r}$ are independently distributed as

$$\hat{\mathbb{y}} \sim \mathcal{N}\big(\mathbf{A}\mathfrak{x}, \sigma^2\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\big)$$

and

$$\mathbb{r} \sim \mathcal{N}\big(0, \sigma^2(\mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T)\big).$$

Linear regression

Let $\delta\hat{\mathbb{y}}$ be a stochastic variable such that

$$\delta\hat{\mathbb{y}} \sim \mathcal{N}\big(0, \tau^2 \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\big).$$

Under the Hypotheses of Gauss-Markov and assuming that $\hat{\mathbb{y}}$ and $\delta\hat{\mathbb{y}}$ are independently distributed, we have

$$\hat{\mathbb{y}} + \delta\hat{\mathbb{y}} \sim \mathcal{N}\big(\mathbf{A}\mathfrak{x}, (\tau^2 + \sigma^2)\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\big).$$

Moreover, we have that

$$||\delta\hat{\mathbb{y}}||_2^2 \sim \tau^2\chi^2(n).$$

Linear regression: a perturbation Theorem

Let $\delta\hat{\mathbb{y}}$ be a stochastic variable such that

$$\delta\hat{\mathbb{y}} \sim \mathcal{N}\big(0, \tau^2 \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\big).$$

Under the hypotheses of Gauss-Markov Theorem and assuming that $\hat{\mathbb{y}}$ and $\delta\hat{\mathbb{y}}$ are uncorrelated, there exist two stochastic variables

$$\begin{aligned}&\delta\mathbb{x}^* \sim \mathcal{N}(0, \tau^2(\mathbf{A}^T\mathbf{A})^{-1}),\\ &\delta\mathbb{y} \sim \mathcal{N}(0, \tau^2\mathbf{I}_m),\end{aligned}$$

such that

1. $\hat{\mathbb{y}} + \delta\hat{\mathbb{y}} = \mathbf{A}(\mathbb{x}^* + \delta\mathbb{x}^*)$,
2. $\mathbb{x}^* + \delta\mathbb{x}^*$ is MVU estimator of $\mathfrak{x}$ for

$$\mathbb{y} + \delta\mathbb{y} = \mathbf{A}\mathfrak{x} + \bar{\mathbb{e}}, \quad \bar{\mathbb{e}} \sim \mathcal{N}(0, (\sigma^2 + \tau^2)\mathbf{I}_m),$$

3. and

$$\bar{\mathbb{s}}^2 = \frac{1}{m-n}||\mathbb{y} + \delta\mathbb{y} - \mathbf{A}(\mathbb{x}^* + \delta\mathbb{x}^*)||_2^2,$$

is the estimator for $\rho^2 = \sigma^2 + \tau^2$ with $\bar{\mathbb{s}}^2 \sim \frac{\sigma^2+\tau^2}{m-n}\chi^2(m-n)$.

Least squares problem

The minimum-variance unbiased (MVU) estimators of $\mathfrak{x}$ and σ^2 are closely related to the solution of the least-squares problem (LSP),

$$\min_{\mathbf{x}} ||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2$$

where $\mathbf{y}$ is a realization of $\mathbb{y}$. The least-squares problem (LSP) has the unique solution

$$\mathbf{x}^* = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y},$$

and the corresponding minimum value is achieved by the square of the euclidean norm of

$$\mathbf{r} = \mathbf{y} - \mathbf{A}\mathbf{x}^* = (\mathbf{I} - \mathbf{P})\mathbf{y}$$

where the matrix $\mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is the orthogonal projector onto $\mathrm{Ker}(\mathbf{A}^T)$ and $\mathbf{P}$ is the orthogonal projector onto $\mathrm{Range}(\mathbf{A})$.

Least squares problem

We remark here that the solution of LSP is deterministic and, therefore, supplies only a realization of the MVU $\mathbb{x}^*$ and of $\mathbb{s}^2$ the corresponding estimator of σ^2.
The vector $\mathbf{x}^*$ is also the solution of the normal equations, i.e. it is the unique stationary point of $||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2$:

$$\mathbf{A}^T\mathbf{A}\mathbf{x}^* = \mathbf{A}^T\mathbf{y}.$$

We will denote its residual in the following by

$$R(\mathbf{x}) = \mathbf{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x})$$

Least squares problem

Given a vector $\tilde{\mathbf{x}} \in \mathrm{R}^n$, the following relations are satisfied:

$$\begin{array}{rcl} \big(\mathbf{I} - \mathbf{P}\big)\ \big(\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}\big) & = & \big(\mathbf{I} - \mathbf{P}\big)\mathbf{y}, \\ \big(\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}\big) & = & \big(\mathbf{y} - \mathbf{A}\mathbf{x}^*\big) + \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\big(\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}\big) \\ & = & \big(\mathbf{y} - \mathbf{A}\mathbf{x}^*\big) + \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}R(\tilde{\mathbf{x}}), \end{array}$$

and, then, we have

$$||\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}||_2^2 = ||\mathbf{y} - \mathbf{A}\mathbf{x}^*||_2^2 + ||R(\tilde{\mathbf{x}})||^2_{(\mathbf{A}^T\mathbf{A})^{-1}},$$

owing to the orthogonality between $\mathbf{y} - \mathbf{A}\mathbf{x}^*$ and $\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}R(\tilde{\mathbf{x}})$.

Least squares problem

From the orthogonality of the projector $\mathbf{P}$, the following are satisfied

$$\begin{array}{rcl} \mathbf{y} & = & \mathbf{P}\mathbf{y} + \big(\mathbf{I} - \mathbf{P}\big)\mathbf{y}, \\ ||\mathbf{y}||_2^2 & = & ||\mathbf{P}\mathbf{y}||_2^2 + ||\big(\mathbf{I} - \mathbf{P}\big)\mathbf{y}||_2^2, \\ ||\mathbf{y}||_2^2 - ||\mathbf{P}\mathbf{y}||_2^2 & = & ||\big(\mathbf{I} - \mathbf{P}\big)\mathbf{y}||_2^2 = ||\mathbf{y} - \mathbf{A}\mathbf{x}^*||_2^2. \end{array}$$

Moreover, we have

$$||\mathbf{P}\mathbf{y}||_2^2 = \mathbf{y}^T\mathbf{A}\big(\mathbf{A}^T\mathbf{A}\big)^{-1}\mathbf{A}^T\mathbf{y} = \mathbf{x}^{*T}\mathbf{A}^T\mathbf{A}\mathbf{x}^*,$$

and, then we conclude that

$$||\mathbf{y}||_2^2 - ||\mathbf{x}^*||_{\mathbf{A}^T\mathbf{A}}^2 = ||\big(\mathbf{I} - \mathbf{P}\big)\mathbf{y}||_2^2 = ||\mathbf{y} - \mathbf{A}\mathbf{x}^*||_2^2.$$

Least squares problem

Finally, it is easy to verify that, given $\tilde{\mathbf{x}}$ as an approximation of $\mathbf{x}^*$,

$$\delta\mathbf{y} = -\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}R(\tilde{\mathbf{x}})$$

is the minimum norm solution of

$$\min_{\mathbf{w}} ||\mathbf{w}||_2^2 \qquad \text{such that} \qquad \mathbf{A}^T\mathbf{A}\tilde{\mathbf{x}} = \mathbf{A}^T(\mathbf{y}+\mathbf{w}).$$

Moreover, using $R(\tilde{\mathbf{x}}) = \mathbf{A}^T(\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}) = \mathbf{A}^T\mathbf{A}(\mathbf{x}^* - \tilde{\mathbf{x}})$, we have

$$||\delta\mathbf{y}||_2^2 = ||R(\tilde{\mathbf{x}})||_{(\mathbf{A}^T\mathbf{A})^{-1}}^2 = ||\mathbf{x}^* - \tilde{\mathbf{x}}||_{\mathbf{A}^T\mathbf{A}}^2.$$

Probabilistic tests and perturbation theory

We have that expression

$$||\delta \mathbf{y}||_2^2 = ||R(\tilde{\mathbf{x}})||^2_{(\mathbf{A}^T\mathbf{A})^{-1}} = ||\mathbf{x}^* - \tilde{\mathbf{x}}||^2_{\mathbf{A}^T\mathbf{A}}.$$

gives a useful key to understand our stopping criteria and their probabilistic nature. If $\mathbf{y}$ can be seen as a realization of a stochastic variable

$$\delta\hat{\mathbb{y}} \sim \mathcal{N}\left(0, \tau^2\mathbf{P}\right)$$

then, based on the perturbation Theorem, the values $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{r}} = \mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}$ are realizations of the stochastic variables associated to

$$\mathbb{y} + \delta\mathbb{y} = \mathbf{A}\mathbb{x} + \bar{\mathbb{e}}, \quad \bar{\mathbb{e}} \sim \mathcal{N}(0, (\sigma^2 + \tau^2)\mathbf{I}_m),$$

Probabilistic tests and perturbation theory

We have that expression

$$||\delta \mathbf{y}||_2^2 = ||R(\tilde{\mathbf{x}})||^2_{(\mathbf{A}^T\mathbf{A})^{-1}} = ||\mathbf{x}^* - \tilde{\mathbf{x}}||^2_{\mathbf{A}^T\mathbf{A}}.$$

gives a useful key to understand our stopping criteria and their probabilistic nature. If $\mathbf{y}$ can be seen as a realization of a stochastic variable

$$\delta\hat{\mathbb{y}} \sim \mathcal{N}\left(0, \tau^2\mathbf{P}\right)$$

then, based on the perturbation Theorem, the values $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{r}} = \mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}$ are realizations of the stochastic variables associated to

$$\mathbb{y} + \delta\mathbb{y} = \mathbf{A}\mathbb{x} + \bar{\mathbb{e}}, \quad \bar{\mathbb{e}} \sim \mathcal{N}(0, (\sigma^2 + \tau^2)\mathbf{I}_m),$$

In practice, we can only check the plausibility of this hypothesis using statistical tests. Fixing some probability threshold η, we check if there is any statistical reason for refusing the previous hypothesis, i.e. the probability we are wrong is very low ($< \eta$).

χ^2 distribution test

$\delta\mathbf{y}$ is a projection onto $\mathrm{Range}(\mathbf{A})$. If $\delta\mathbf{y}$ is a realization of a stochastic variable $\delta\hat{\mathbb{y}}$ satisfying

$$\min_{\mathbf{w}} ||\mathbf{w}||_2^2 \qquad \text{such that} \qquad \mathbf{A}^T\mathbf{A}\tilde{\mathbf{x}} = \mathbf{A}^T(\mathbf{y}+\mathbf{w}).$$

then $\|\delta y\|_2^2$ is a realization of $\|\delta\hat{\mathbb{y}}\|_2^2 \sim \tau^2\chi^2(n)$.

χ^2 distribution test

Therefore, we consider that δy is a sample of the stochastic variable $\delta\hat{\mathbb{y}}$, if for some small enough η,

$$\mathrm{Probability}(\|\delta\hat{\mathbb{y}}\|_2^2 \geq \|\delta y\|_2^2) \geq 1 - \eta,$$

where we assume that the random variable $\frac{\|\delta\hat{\mathbb{y}}\|_2^2}{\tau^2}$ follows a centered χ^2 distribution with n degrees of freedom.

χ^2 distribution test

Thus, we can formulate our criterion as

$$p_\chi\left(\frac{\|\delta\mathbf{y}\|_2^2}{\tau^2}, n\right) \equiv \mathrm{Probability}\left(\frac{\|\delta\hat{\mathbb{y}}\|_2^2}{\tau^2} \leq \frac{\|\delta\mathbf{y}\|_2^2}{\tau^2}\right) \leq \eta, \qquad (5)$$

where, since $\delta\hat{\mathbb{y}}$ is a Gaussian distribution with covariance matrix $\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$, the value of $p_\chi(., n)$ is the cumulative distribution function of the χ^2 distribution Abramowitz-Stegun (26.4): The probability that $\mathbb{X}^2 = \sum_i \mathbb{X}_i^2$ with ν degrees of freedom $\mathbb{X}_i \sim \mathcal{N}(0,1)$, is such that $X^2 \leq \chi^2$ is

$$Prob(\chi^2|\nu) = \left[2^{\nu/2}\Gamma(\frac{\nu}{2})\right]^{-1} \int_0^{\chi^2} t^{\frac{\nu}{2}-1} e^{\frac{t}{2}} dt.$$

χ^2 distribution test

Moreover, the corresponding $\tilde{\mathbf{x}}$ is the exact solution of

$$\mathbf{A}^T\mathbf{A}\tilde{\mathbf{x}} = \mathbf{A}^T(\mathbf{y} + \delta\mathbf{y}).$$

Thus, we can interpret $\tilde{\mathbf{x}}$ as a realization of the stochastic variable $\mathbb{x}^* + \delta\mathbb{x}^*$ and $\mathbf{A}\tilde{\mathbf{x}}$ as a realization of $\mathbb{y} + \delta\mathbb{y}$: i.e. we have, with probability η, realizations consistent with the perturbed linear regression problem, that, if we choose $\tau^2 \ll \sigma^2$, is only marginally different from the original.

Stopping criteria for CGLS

If we use the conjugate gradient method in order to compute the solution, it is quite natural to have a stopping criterion which takes advantage of the minimization property of this method and of the stochastic properties of the underpinning problem

Stopping criteria for CGLS

PCGLS algorithm Given an initial guess $x^{(0)}$, compute $r^{(0)} = \left(y - Ax^{(0)}\right)$, $R^{(0)} = A^T r^{(0)}$, and solve $Mz^{(0)} = R^{(0)}$. Set $q^{(0)} = z^{(0)}$, $\beta_0 = 0$, $\nu_0 = 0$, $\chi_1 = R^{(0)T} z^{(0)}$, and $\xi_{-d} = \infty$.

```
$k = 0$
while $\mathbf{z}(\xi_{k-d}, \|r^{(0)}\|_2, \nu_k, \tau^2, \sigma^2)) > \eta$ do
    $k = k + 1$;
    $p = Aq^{(k-1)}$;
    $\alpha_{k-1} = \chi_k / ||p||_2^2$;
    $\psi_k = \alpha_{k-1}\chi_k$; $\nu_k = \nu_{k-1} + \psi_k$;
    $x^{(k)} = x^{(k-1)} + \alpha_{k-1} q^{(k-1)}$;
    $R^{(k)} = R^{(k-1)} - \alpha_{k-1} A^T q^{(k-1)}$;
    Solve $Mz^{(k)} = R^{(k)}$;
    $\chi_{k+1} = R^{(k)T} z^{(k)}$ ;
    $\beta_k = \chi_{k+1}/\chi_k$;
    $q^{(k)} = z^{(k)} + \beta_k q^{(k-1)}$;
    if $k > d$ then
        $\xi_{k-d} = \sum_{j=k-d+1}^{k} \psi_j$;
    else
        $\xi_{k-d} = \infty$;
    endif
end while.
```

χ^2 stopping criteria for PCGLS

To detect the convergence as early as possible and avoid over-solving in the LSP, we consider a $\delta\mathbf{y}_k$ with minimum Euclidean norm such that $\mathbf{x}^{(k)}$ exactly solves a LS problem. Using the estimations we have

$$\text{IF} \quad p_\chi\left(\frac{\xi_k}{\tau^2}, n\right) \leq \eta \quad \text{THEN STOP} .$$

In order to have perturbations of $\hat{\mathbf{y}}$ that not distort excessively the statistical properties of the original linear regression, we assume that $\tau^2 \ll \sigma^2$.

χ^2 stopping criteria for PCGLS

We re-iterate that χ^2 test is a measure of the probability that the numerical values computed at step k will be statistically equivalent to those obtained solving an LSP related to a perturbed linear regression model exactly where the statistical errors $\mathbf{e} \sim \mathcal{N}\big(0, (\sigma^2 + \tau^2) I_m\big)$, i.e. small value of η will indicate that the probability of stopping at the wrong place is small.

χ^2 stopping criteria for PCGLS

We re-iterate that χ^2 test is a measure of the probability that the numerical values computed at step k will be statistically equivalent to those obtained solving an LSP related to a perturbed linear regression model exactly where the statistical errors $\mathbf{e} \sim \mathcal{N}\left(0, (\sigma^2 + \tau^2) I_m\right)$, i.e. small value of η will indicate that the probability of stopping at the wrong place is small. In PCGLS, we can choose

$$\mathbf{z}(\xi_k, \|r^{(0)}\|_2, \nu_k, \tau^2, \sigma^2) = p_\chi\left(\frac{\xi_k(m-n)}{\tau^2}, n\right).$$

Choice of η and τ^2 for the χ^2 and **F**-test stopping criteria

We seek choices of η and τ^2 that will depend on the properties of the problem that we want to solve and, in the practical cases, we would like η and τ^2 to be much larger than ε, the roundoff unit of the computer finite precision arithmetic.

Choice of η and τ^2 for the χ^2 and **F**-test stopping criteria

The choice of η is related to the probability the user subjectively feels as adequate, i.e. he/she accepts that the probability of choosing the wrong iterate is less than η. In our experiments, we chose $\eta = 10^{-8}$ which is quite conservative. This value is close to the probability of winning the lotto.

Choice of η and τ^2 for the χ^2 and **F**-test stopping criteria

The choice of τ^2 is also related to a priori knowledge of the statistical properties of the linear regression problem and in particular to the user knowledge of a reliable value of σ^2 or of an interval where σ^2 lies. We experimented with several values of τ^2. The numerical results suggest that the choice $\tau^2 = \sigma^2$ gives reliable answers in the majority of our tests and they are always consistent with the results of Theorem on perturbations. When σ^2 is approximated by its upper bound $(\|y\|_2^2 - \nu_k)/(m-n)$ and the dynamical choices are used $\tau_k^2 = (\|y\|_2^2 - \nu_k)/(m-n)$ we can have an early stop because $(\|y\|_2^2 - \nu_k)/(m-n)$ is a poor approximation of the true standard deviation. However, smaller values of τ_k^2 ($\tau_k^2 = 0.1(\|y\|_2^2 - \nu_k)/(m-n)$ or $\tau_k^2 = 0.01(\|y\|_2^2 - \nu_k)/(m-n)$) proved more robust and reliable. In these cases it would be useful to have lower bound approximations of σ^2. Unfortunately, to compute a lower bound of $(\|y\|_2^2 - \nu_k)/(m-n)$ can be costly.

Choice of η and τ^2 for the χ^2 and **F**-test stopping criteria

The values of ξ_k and ν_k are lower bounds respectively for the true energy norm of the errors and the energy norm of the solution, which are both independent of the preconditioner used. However, a good preconditioner will help to reduce the delay factor d. i.e. we will have better approximations at a cheaper computational cost.

Data assimilation test

Data assimilation problems constitute an important class of regression problems. Their purpose is to reconstruct the initial conditions at $t = 0$ of a dynamical system based on knowledge of the system's evolution laws and on observations of the state at times t_i. More precisely, consider a linear dynamical system described by the equation $\dot{u} = f(t, u)$ whose solution operator is given by $u(t) = M(t)u_0$. Assume that the system state is observed (possibly only in parts) at times $\{t_i\}_{i=0}^{N}$, yielding observation vectors $\{y_i\}_{i=0}^{N}$, whose model is given by $y_i = Hu(t_i) + \epsilon$, where ϵ is a noise with covariance matrix $R_i = \sigma^2 I$.

Data assimilation test

We are then interested in finding u_0 which minimizes

$$\frac{1}{2}\sum_{i=0}^{N}\|HM(t_i)u_0 - y_i\|^2_{R_i^{-1}}.$$

We consider here the case where the dynamical system is the linear heat equation in a two-dimensional domain, defined on $S_2 = [0,1]\times[0,1]$ by

$$\frac{\partial u}{\partial t} = -\Delta u \quad in \quad S_2, \quad u = 0 \quad on \quad \partial S_2, \quad u(.,0) = u_0 \quad in \quad S_2.$$

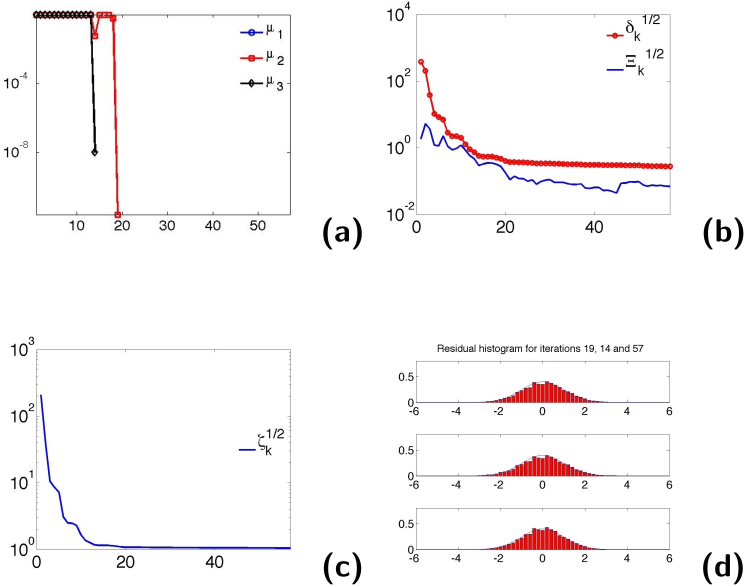
Data assimilation test

The system is integrated with timestep dt, using an implicit Euler scheme. In the physical domain, a regular finite difference scheme is taken for the Laplace operator, with same spacing h in the two spatial dimensions. The data of our problem is computed by imposing a solution $u_0(x, y, 0)$ computing the exact system trajectory and observing Hu at every point in the spatial domain and at every time step. In our application, $m = 8100$, $n = 900 = 30^2$, $dt = 1$, $h = 1/31$, $N = 8$ and $H = \mathrm{diag}(1^{1.5}, 2^{1.5}, \ldots, n^{1.5})$. The observation vector y is obtained by imposing $u_0(x, y, 0) = \frac{1}{4}\sin(\frac{1}{4}x)(x-1)\sin(5y)(y-1)$, and by adding a random measurement error with Gaussian distribution with zero mean and covariance matrix $R_i = \sigma^2 I_n$, where $\sigma = 10^{-3}$. In our numerical experiments, we use PCGLS without preconditioner.

Data Assimilation test: results

Our choice of not using a preconditioner is not optimal, however, the choice of $d = 5$ in this problem gives reliable answers and stable behaviour of the stopping criteria.

Data Assimilation test: results



Lecture on $\mathbf{LDL}^T$ multifrontal and GMRES and FGMRES

Outline

- GMRES and Flexible GMRES
- Multifrontal
- Static pivoting
- Roundoff error analysis
- Mixed precision
- Test problems
- Numerical experiments

GMRES and FGMRES

Let $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ and $\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$ be the usual Krylov space
GMRES

$$\min_{\mathbf{x}\in\mathbf{x}_0+\mathcal{K}_k(\mathbf{A},\mathbf{r}_0)} ||\mathbf{r}_0 - \mathbf{A}\mathbf{x}||_2 \qquad \mathbf{r}_0 - \mathbf{A}\mathbf{x}_k \perp \mathbf{A}\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$$

GMRES and FGMRES

Let $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ and $\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$ be the usual Krylov space GMRES

$$\min_{\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)} ||\mathbf{r}_0 - \mathbf{A}\mathbf{x}||_2 \qquad \mathbf{r}_0 - \mathbf{A}\mathbf{x}_k \perp \mathbf{A}\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$$

GMRES Left preconditioning

$$\mathbf{L}^{-1}\mathbf{A}\mathbf{x} = \mathbf{L}^{-1}\mathbf{b} \quad \left\{ \begin{array}{l} (\mathbf{L}^{-1}\mathbf{A}, \mathbf{L}^{-1}\mathbf{b}) \longrightarrow (\mathbf{A}, \mathbf{b}) \\ \mathcal{K}_k(\mathbf{L}^{-1}\mathbf{A}, \mathbf{L}^{-1}\mathbf{r}_0) \longrightarrow \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0) \end{array} \right.$$

changes the norm.

GMRES and FGMRES

Let $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ and $\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$ be the usual Krylov space
GMRES

$$\min_{\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)} ||\mathbf{r}_0 - \mathbf{A}\mathbf{x}||_2 \qquad \mathbf{r}_0 - \mathbf{A}\mathbf{x}_k \perp \mathbf{A}\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$$

GMRES Right preconditioning

$$\mathbf{A}\mathbf{M}^{-1}\mathbf{y} = \mathbf{b} \quad \left\{ \begin{array}{l} (\mathbf{A}\mathbf{M}^{-1}, \mathbf{r}_0) \longrightarrow (\mathbf{A}, \mathbf{r}_0) \\ \mathcal{K}_k(\mathbf{A}\mathbf{M}^{-1}, \mathbf{r}_0) \longrightarrow \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0) \\ \mathbf{x}_k = \mathbf{M}^{-1}\mathbf{y}_k \\ \mathbf{A}\mathbf{M}^{-1}\mathbf{V}_k = \mathbf{V}_{k+1}\mathbf{H}_k \end{array} \right.$$

GMRES and FGMRES

Let $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ and $\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$ be the usual Krylov space
GMRES

$$\min_{\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)} ||\mathbf{r}_0 - \mathbf{A}\mathbf{x}||_2 \qquad \mathbf{r}_0 - \mathbf{A}\mathbf{x}_k \perp \mathbf{A}\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$$

GMRES Right preconditioning

$$\mathbf{A}\mathbf{M}^{-1}\mathbf{y} = \mathbf{b} \quad \begin{cases} (\mathbf{A}\mathbf{M}^{-1}, \mathbf{r}_0) \longrightarrow (\mathbf{A}, \mathbf{r}_0) \\ \mathcal{K}_k(\mathbf{A}\mathbf{M}^{-1}, \mathbf{r}_0) \longrightarrow \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0) \\ \mathbf{x}_k = \mathbf{M}^{-1}\mathbf{y}_k \\ \mathbf{A}\mathbf{M}^{-1}\mathbf{V}_k = \mathbf{V}_{k+1}\mathbf{H}_k \end{cases}$$

Flexible GMRES Right preconditioning

$$\mathbf{Z}_k \longrightarrow \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0), \quad \mathbf{x}_k = \mathbf{x}_0 + \mathbf{Z}_k\mathbf{y}_k \quad \mathbf{A}\mathbf{Z}_k = \mathbf{V}_{k+1}\mathbf{H}_k$$

$$\mathbf{Z}_k = span(\mathbf{r}_0, \mathbf{A}\mathbf{M}_1^{-1}\mathbf{r}_0, \ldots, \left(\prod_{j=0}^{k-1} \mathbf{A}\mathbf{M}_j^{-1}\right) \mathbf{r}_0)$$

Linear system

We wish to solve large sparse systems

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

where $\mathbf{A} \in \mathrm{R}^{\mathbf{N}\times\mathbf{N}}$ is symmetric indefinite

Linear system

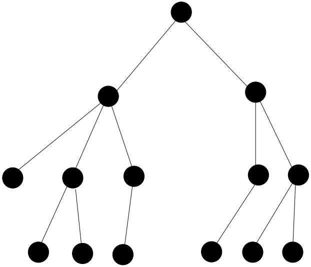
A particular and important case arises in saddle-point problems where the coefficient matrix is of the form

$$\begin{bmatrix} \mathbf{H} & \mathbf{A} \\ \mathbf{A}^T & 0 \end{bmatrix}$$

Since we want accurate solutions and norm-wise backward stability, we will use as preconditioners fast factorizations of **A** computed using static pivoting or mixed precision arithmetic.

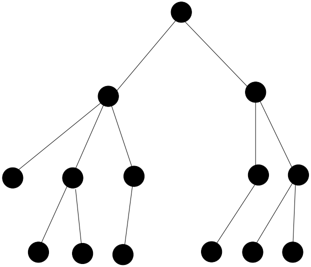
Multifrontal method

ASSEMBLY TREE



Multifrontal method

ASSEMBLY TREE

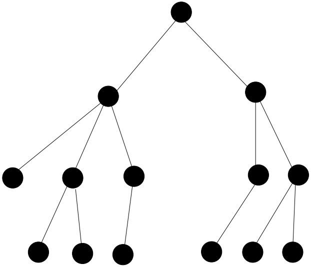


AT EACH NODE

$$\begin{array}{|c|c|} \hline F_{11} & F_{12} \\ \hline F_{12}^{T} & F_{22} \\ \hline \end{array}$$

Multifrontal method

ASSEMBLY TREE

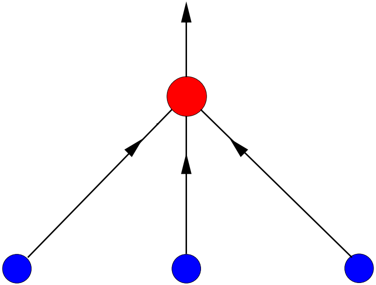


AT EACH NODE

$$\begin{array}{|c|c|}\hline F_{11} & F_{12} \\ \hline F_{12}^T & F_{22} \\ \hline\end{array}$$

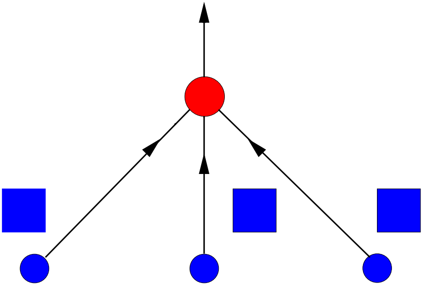
$$F_{22} \leftarrow F_{22} \; - \; F_{12}^T F_{11}^{-1} F_{12}$$

Multifrontal method



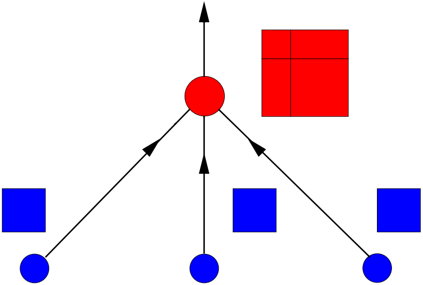
- From children to parent

Multifrontal method



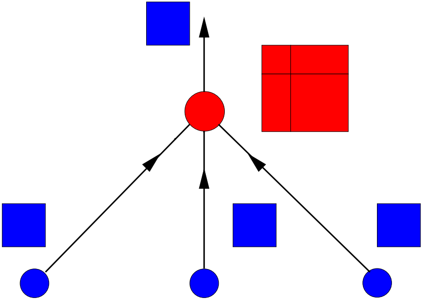
- From children to parent
- **ASSEMBLY**
  Gather/Scatter operations
  (indirect addressing)

Multifrontal method



- From children to parent
- **ASSEMBLY**
  Gather/Scatter operations (indirect addressing)
- **ELIMINATION** Full Gaussian elimination, Level 3 BLAS (TRSM, GEMM)

Multifrontal method



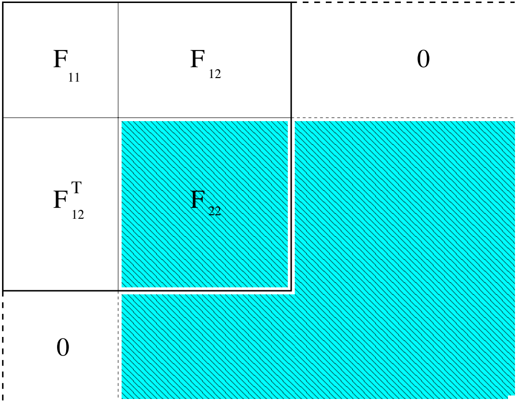
- From children to parent
- **ASSEMBLY**
  Gather/Scatter operations (indirect addressing)
- **ELIMINATION** Full Gaussian elimination, Level 3 BLAS (TRSM, GEMM)

Multifrontal method

$$\left[\begin{array}{c|c} F_{11} & F_{12} \\ \hline F_{12}^{T} & F_{22} \end{array}\right]$$

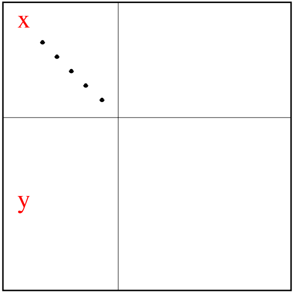
Pivot can only be chosen fromF_{11} block since F_{22} is **NOT** fully summed.

Multifrontal method



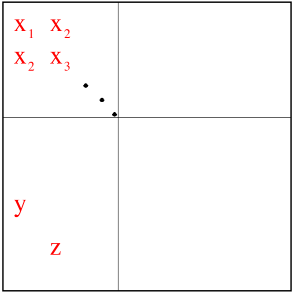
Situation wrt rest of matrix

Pivoting (1×1)



Choose x as 1×1 **pivot** if $|x| > u|y|$
where $|y|$ is the largest in column.

Pivoting (2 × 2)

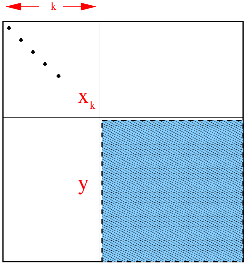


For the indefinite case, we can choose 2×2 **pivot** where we require

$$\left| \begin{bmatrix} x_1 & x_2 \\ x_2 & x_3 \end{bmatrix}^{-1} \right| \begin{bmatrix} |y| \\ |z| \end{bmatrix} \leq \begin{bmatrix} \frac{1}{u} \\ \frac{1}{u} \end{bmatrix}$$

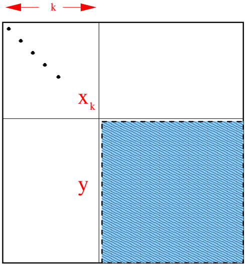
where again $|y|$ and $|z|$ are the largest in their columns.

Pivoting



If we assume that $k-1$ pivots are chosen but $|x_k| < u|y|$:

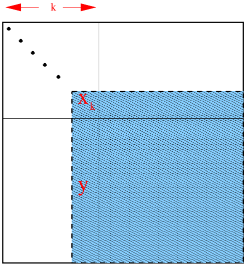
Pivoting



If we assume that $k-1$ pivots are chosen but $|x_k| < u|y|$:

- we can either take the **RISK** and use it or

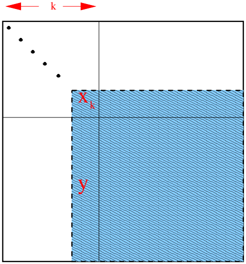
Pivoting



If we assume that $k-1$ pivots are chosen but $|x_k| < u|y|$:

- we can either take the **RISK** and use it or
- **DELAY** the pivot and then send to the parent a larger Schur complement.

Pivoting



If we assume that $k-1$ pivots are chosen but $|x_k| < u|y|$:

- we can either take the **RISK** and use it or
- **DELAY** the pivot and then send to the parent a larger Schur complement.

This can cause more work and storage

Static Pivoting

An **ALTERNATIVE** is to use **Static Pivoting**, by replacing x_k by

$$x_k + \tau$$

and CONTINUE.

Static Pivoting

An **ALTERNATIVE** is to use **Static Pivoting**, by replacing x_k by

$$x_k + \tau$$

and CONTINUE.

This is even more important in the case of parallel implementation where static data structures are often preferred

Static Pivoting

Several codes use (or have an option for) this device:

- SuperLU (Demmel and Li)
- PARDISO (Gärtner and Schenk)
- MA57 (Duff and Pralet)

Static Pivoting

We thus have factorized

$$\mathbf{A} + \mathbf{E} = \mathbf{L}\mathbf{D}\mathbf{L}^T = \mathbf{M}$$

where $|\mathbf{E}| \leq \tau \mathbf{I}$

Static Pivoting

We thus have factorized

$$\mathbf{A} + \mathbf{E} = \mathbf{LDL}^T = \mathbf{M}$$

where $|\mathbf{E}| \leq \tau \mathbf{I}$

The three codes then have an **Iterative Refinement** option.
IR will converge if $\rho(\mathbf{M}^{-1}\mathbf{E}) < 1$

Static Pivoting

Choosing τ

Static Pivoting

Choosing τ

Increase $\tau \Longrightarrow$ increase stability of decomposition

Static Pivoting

Choosing τ

Increase $\tau \Longrightarrow$ increase stability of decomposition

Decrease $\tau \Longrightarrow$ better approximation of the original matrix, reduces $||\mathbf{E}||$

Static Pivoting

Choosing τ

Increase $\tau \Longrightarrow$ increase stability of decomposition

Decrease $\tau \Longrightarrow$ better approximation of the original matrix, reduces $||\mathbf{E}||$

Trade-off

- $\approx \varepsilon \Longrightarrow$ big growth in preconditioning matrix $\mathbf{M}$
- $\approx 1 \Longrightarrow$ huge error $||\mathbf{E}||$.

Static Pivoting

Choosing τ

Increase $\tau \Longrightarrow$ increase stability of decomposition

Decrease $\tau \Longrightarrow$ better approximation of the original matrix, reduces $||\mathbf{E}||$

Trade-off

- $\approx \varepsilon \Longrightarrow$ big growth in preconditioning matrix $\mathbf{M}$
- $\approx 1 \Longrightarrow$ huge error $||\mathbf{E}||$.

Conventional wisdom is to choose

$$\tau = \mathcal{O}(\sqrt{\varepsilon})$$

Static Pivoting

Choosing τ

Increase $\tau \implies$ increase stability of decomposition

Decrease $\tau \implies$ better approximation of the original matrix, reduces $||\mathbf{E}||$

Trade-off

- $\approx \varepsilon \implies$ big growth in preconditioning matrix $\mathbf{M}$
- $\approx 1 \implies$ huge error $||\mathbf{E}||$.

Conventional wisdom is to choose

$$\tau = \mathcal{O}(\sqrt{\varepsilon})$$

In real life $\rho(M^{-1}E) > 1$

Right preconditioned GMRES

```
procedure [x] = right_Prec_GMRES(A,M,b)
        $\mathbf{x}_0 = \mathbf{M}^{-1}\mathbf{b}$, $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ and $\beta = ||\mathbf{r}_0||$
        $\mathbf{v}_1 = \mathbf{r}_0/\beta$; k = 0;
        while $||\mathbf{r}_k|| > \mu(||\mathbf{b}|| + ||\mathbf{A}|| \, ||\mathbf{x}_k||)$
            $k = k + 1$;
            $\mathbf{z}_k = \mathbf{M}^{-1}\mathbf{v}_k$; $\mathbf{w} = \mathbf{A}\mathbf{z}_k$;
            for $i = 1, \ldots, k$ do
                $h_{i,k} = \mathbf{v}_i^T\mathbf{w}$ ;
                $\mathbf{w} = \mathbf{w} - h_{i,k}\mathbf{v}_i$;
            end for;
            $h_{k+1,k} = ||\mathbf{w}||$;
            $\mathbf{v}_{k+1} = \mathbf{w}/h_{k+1,k}$;
            $\mathbf{V}_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$; $\mathbf{H}_k = \{h_{i,j}\}_{1\leq i\leq j+1; 1\leq j\leq k}$;
            $\mathbf{y}_k = \arg\min_{\mathbf{y}} ||\beta\mathbf{e}_1 - \mathbf{H}_k\mathbf{y}||$;
            $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{M}^{-1}\mathbf{V}_k\mathbf{y}_k$ and $\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k$;
        end while ;
end procedure.
```

Right preconditioned Flexible GMRES

```
procedure $[\mathbf{x}] = \text{FGMRES}(\mathbf{A},\mathbf{M}_i,\mathbf{b})$
    $\mathbf{x}_0 = \mathbf{M}_0^{-1}\mathbf{b}$, $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ and $\beta = ||\mathbf{r}_0||$
    $\mathbf{v}_1 = \mathbf{r}_0/\beta$; $\mathrm{k} = 0$;
    while $||\mathbf{r}_k|| > \mu(||\mathbf{b}|| + ||\mathbf{A}||\ ||\mathbf{x}_k||)$
        $k = k + 1$;
        $\mathbf{z}_k = \mathbf{M}_k^{-1}\mathbf{v}_k$; $\mathbf{w} = \mathbf{A}\mathbf{z}_k$;
        for $i = 1, \ldots, k$ do
            $h_{i,k} = \mathbf{v}_i^T\mathbf{w}$ ;
            $\mathbf{w} = \mathbf{w} - h_{i,k}\mathbf{v}_i$;
        end for;
        $h_{k+1,k} = ||\mathbf{w}||$; $\mathbf{v}_{k+1} = \mathbf{w}/h_{k+1,k}$;
        $\mathbf{Z}_k = [\mathbf{z}_1, \ldots, \mathbf{z}_k]$; $\mathbf{V}_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$;
        $\mathbf{H}_k = \{h_{i,j}\}_{1\le i\le j+1; 1\le j\le k}$;
        $\mathbf{y}_k = \arg\min_{\mathbf{y}} ||\beta\mathbf{e}_1 - \mathbf{H}_k\mathbf{y}||$;
        $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{Z}_k\mathbf{y}_k$ and $\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k$;
    end while ;
end procedure.
```

Roundoff error 1

The computed $\hat{\mathbf{L}}$ and $\hat{\mathbf{D}}$ in floating-point arithmetic satisfy

$$\begin{cases} \mathbf{A} + \delta\mathbf{A} + \tau\mathbf{E} = \mathbf{M} \\ ||\delta\mathbf{A}|| \leq c(n)\varepsilon\, ||\,|\hat{\mathbf{L}}|\,|\hat{\mathbf{D}}|\,|\hat{\mathbf{L}}^T|\,|| \\ ||\mathbf{E}|| \leq 1. \end{cases}$$

The perturbation $\delta\mathbf{A}$ must have a norm smaller than τ, in order to not dominate the global error.

Roundoff error 1

The computed $\hat{\mathbf{L}}$ and $\hat{\mathbf{D}}$ in floating-point arithmetic satisfy

$$\begin{cases} \mathbf{A} + \delta\mathbf{A} + \tau\mathbf{E} = \mathbf{M} \\ ||\delta\mathbf{A}|| \leq c(n)\varepsilon \, || \, |\hat{\mathbf{L}}| \, |\hat{\mathbf{D}}| \, |\hat{\mathbf{L}}^T| \, || \\ ||\mathbf{E}|| \leq 1. \end{cases}$$

The perturbation $\delta\mathbf{A}$ must have a norm smaller than τ, in order to not dominate the global error.

A sufficient condition for this is $\boxed{n \, \varepsilon \, || \, |\hat{\mathbf{L}}| \, |\hat{\mathbf{D}}| \, |\hat{\mathbf{L}}^T| \, || \leq \tau}$

Roundoff error 1

The computed $\hat{\mathbf{L}}$ and $\hat{\mathbf{D}}$ in floating-point arithmetic satisfy

$$\begin{cases} \mathbf{A} + \delta\mathbf{A} + \tau\mathbf{E} = \mathbf{M} \\ \|\delta\mathbf{A}\| \le c(n)\varepsilon \, \| \, |\hat{\mathbf{L}}| \, |\hat{\mathbf{D}}| \, |\hat{\mathbf{L}}^T| \, \| \\ \|\mathbf{E}\| \le 1. \end{cases}$$

The perturbation $\delta\mathbf{A}$ must have a norm smaller than τ, in order to not dominate the global error.

A sufficient condition for this is $\boxed{n\,\varepsilon \, \| \, |\hat{\mathbf{L}}| \, |\hat{\mathbf{D}}| \, |\hat{\mathbf{L}}^T| \, \| \le \tau}$

$\| \, |\hat{\mathbf{L}}| \, |\hat{\mathbf{D}}| \, |\hat{\mathbf{L}}^T| \, \| \approx \frac{n}{\tau} \Longrightarrow \varepsilon \le \frac{\tau^2}{n^2}$

Roundoff error 1

The computed $\hat{\mathbf{L}}$ and $\hat{\mathbf{D}}$ in floating-point arithmetic satisfy

$$\begin{cases} \mathbf{A} + \delta\mathbf{A} + \tau\mathbf{E} = \mathbf{M} \\ ||\delta\mathbf{A}|| \leq c(n)\varepsilon\, ||\,|\hat{\mathbf{L}}|\,|\hat{\mathbf{D}}|\,|\hat{\mathbf{L}}^T|\,|| \\ ||\mathbf{E}|| \leq 1. \end{cases}$$

The perturbation $\delta\mathbf{A}$ must have a norm smaller than τ, in order to not dominate the global error.

A sufficient condition for this is $\boxed{n\,\varepsilon\,||\,|\hat{\mathbf{L}}|\,|\hat{\mathbf{D}}|\,|\hat{\mathbf{L}}^T|\,|| \leq \tau}$

$||\,|\hat{\mathbf{L}}|\,|\hat{\mathbf{D}}|\,|\hat{\mathbf{L}}^T|\,|| \approx \frac{n}{\tau} \Longrightarrow \varepsilon \leq \frac{\tau^2}{n^2}$

Moreover, we assume that $\boxed{\max\{||\mathbf{M}^{-1}||, ||\bar{\mathbf{Z}}_k||\} \leq \frac{\tilde{c}}{\tau}}$.

Roundoff error

The roundoff error analysis of both FGMRES and GMRES can be made in four stages:

Roundoff error

The roundoff error analysis of both FGMRES and GMRES can be made in four stages:

1. Error analysis of the Arnoldi-Krylov process (Giraud and Langou, Björck and Paige, and Paige, Rozložník, and Strakoš).

   MGS applied to

$$\mathbf{z}_1 = \mathbf{M}_1^{-1}\mathbf{r}_0/||\mathbf{r}_0||, \quad \mathbf{z}_j = \mathbf{M}_j^{-1}\mathbf{v}_j$$
$$\mathbf{C} = (\mathbf{z}_1, \mathbf{A}\mathbf{z}_1, \mathbf{A}\mathbf{z}_2, \dots) = \mathbf{V}_{k+1}\mathbf{R}_k$$

$$\mathbf{R}_k = \left[\begin{array}{cccc} ||\mathbf{r}_0|| & \mathbf{H}_{1,1} & \dots & \mathbf{H}_{1,k} \\ 0 & \mathbf{H}_{2,1} & \dots & \mathbf{H}_{2,k} \\ 0 & 0 & \dots & \mathbf{H}_{3,k} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \mathbf{H}_{k+1,k} \end{array}\right]$$

Roundoff error

The roundoff error analysis of both FGMRES and GMRES can be made in four stages:

1. Error analysis of the Arnoldi-Krylov process (Giraud and Langou, Björck and Paige, and Paige, Rozložník, and Strakoš).
2. Error analysis of the Givens process used on the upper Hessenberg matrix $\mathbf{H}_k$ in order to reduce it to upper triangular form.

Roundoff error

The roundoff error analysis of both FGMRES and GMRES can be made in four stages:

1. Error analysis of the Arnoldi-Krylov process (Giraud and Langou, Björck and Paige, and Paige, Rozložník, and Strakoš).
2. Error analysis of the Givens process used on the upper Hessenberg matrix $\mathbf{H}_k$ in order to reduce it to upper triangular form.
3. Error analysis of the computation of $\mathbf{x}_k$ in FGMRES and GMRES.

Roundoff error

The roundoff error analysis of both FGMRES and GMRES can be made in four stages:

1. Error analysis of the Arnoldi-Krylov process (Giraud and Langou, Björck and Paige, and Paige, Rozložník, and Strakoš).
2. Error analysis of the Givens process used on the upper Hessenberg matrix $\mathbf{H}_k$ in order to reduce it to upper triangular form.
3. Error analysis of the computation of $\mathbf{x}_k$ in FGMRES and GMRES.
4. Use of the static pivoting properties and $\mathbf{A} + \mathbf{E} = \mathbf{LDL}^T$ in order to have the final expressions.

The first two stages of the roundoff error analysis are the same for both FGMRES and GMRES. The last stage is specific to each one of the two algorithms.

Roundoff error FGMRES

Theorem 1.

$$\boxed{\sigma_{min}(\bar{\mathbf{H}}_k) > c_7(k,1)\varepsilon\,||\bar{\mathbf{H}}_k|| + \mathcal{O}(\varepsilon^2) \quad \forall k,}$$

$$\boxed{|\bar{s}_k| < 1 - \varepsilon\,,\ \forall k,}$$

(where $\bar{s}_k$ are the sines computed during the Givens algorithm)

and

$$\boxed{2.12(n+1)\varepsilon\ < 0.01 \textit{ and } 18.53\varepsilon\, n^{\frac{3}{2}}\kappa(\mathbf{C}^{(k)}) < 0.1\ \forall k}$$

$$\exists\hat{k}, \qquad \hat{k} \le n$$

such that, $\forall k \ge \hat{k}$*, we have*

$$||\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_k|| \le c_1(n,k)\varepsilon\left(||\mathbf{b}|| + ||\mathbf{A}||\,||\bar{\mathbf{x}}_0|| + ||\mathbf{A}||\,||\bar{\mathbf{Z}}_k||\,||\bar{\mathbf{y}}_k||\right) + \mathcal{O}(\varepsilon^2).$$

Roundoff error FGMRES

Moreover, if $\mathbf{M}_i = \mathbf{M}, \forall i$,

$$\boxed{\rho = 1.3\,||\hat{\mathbf{W}}_k|| + c_2(k,1)\varepsilon\,||\mathbf{M}||\,||\bar{\mathbf{Z}}_k|| < 1 \quad \forall k < \hat{k},}$$

where

$$\boxed{\hat{\mathbf{W}}_k = \left[\mathbf{M}\bar{\mathbf{z}}_1 - \bar{\mathbf{v}}_1, \ldots, \mathbf{M}\bar{\mathbf{z}}_k - \bar{\mathbf{v}}_k\right],}$$

we have:

$$||\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_k|| \leq$$

$$c(n,k)\gamma\varepsilon\left(||\mathbf{b}|| + ||\mathbf{A}||\,||\bar{\mathbf{x}}_0|| + ||\mathbf{A}||\,||\bar{\mathbf{Z}}_k||\,||\mathbf{M}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_0)||\right) + \mathcal{O}(\varepsilon^2)$$

$$\gamma = \frac{1.3}{1-\rho}.$$

Roundoff error FGMRES

Theorem 2

Under the Hypotheses of Theorem 1, and

$$\boxed{c(n)\varepsilon \,\|\, |\hat{\mathbf{L}}|\, |\hat{\mathbf{D}}|\, |\hat{\mathbf{L}}^T| \,\|\, < \tau}$$

$$\boxed{c(n,k)\gamma\varepsilon \,||\mathbf{A}||\, ||\bar{\mathbf{Z}}_k|| < 1 \quad \forall k < \hat{k}}$$

$$\boxed{\max\{||\mathbf{M}^{-1}||, ||\bar{\mathbf{Z}}_k||\} \leq \frac{\tilde{c}}{\tau}}$$

we have

Roundoff error FGMRES

Theorem 2

Under the Hypotheses of Theorem 1, and

$$\boxed{c(n)\varepsilon\, ||\, |\hat{\mathbf{L}}|\, |\hat{\mathbf{D}}|\, |\hat{\mathbf{L}}^T|\, ||\, < \tau}$$

$$\boxed{c(n,k)\gamma\varepsilon\, ||\mathbf{A}||\, ||\bar{\mathbf{Z}}_k|| < 1 \quad \forall k < \hat{k}}$$

$$\boxed{\max\{||\mathbf{M}^{-1}||, ||\bar{\mathbf{Z}}_k||\} \le \frac{\tilde{c}}{\tau}}$$

we have

$$||\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_k|| \le 2\mu\varepsilon\, (||\mathbf{b}|| + ||\mathbf{A}||\, (||\bar{\mathbf{x}}_0|| + ||\bar{\mathbf{x}}_k||)) + \mathcal{O}(\varepsilon^2).$$

$$\mu = \frac{c(n,k)}{1 - c(n,k)\varepsilon\, ||\mathbf{A}||\, ||\bar{\mathbf{Z}}_k||}$$

Roundoff error right preconditioned GMRES

Theorem 3
We assume of applying Iterative Refinement for solving $\mathbf{M}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_0) = \bar{\mathbf{V}}_k \bar{\mathbf{y}}_k$ *at last step.*
Under the Hypotheses of Theorem 1 and $\boxed{c(n)\varepsilon\ \kappa(M) < 1}$

$$\exists \hat{k}, \qquad \hat{k} \leq n$$

such that, $\forall k \geq \hat{k}$*, we have*

$$\begin{aligned} ||\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_k|| \leq \quad & c_1(n,k)\varepsilon \Big\{ ||\mathbf{b}|| + ||\mathbf{A}||\, ||\bar{\mathbf{x}}_0|| + \\ & ||\mathbf{A}||\, ||\bar{\mathbf{Z}}_k||\, ||\mathbf{M}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_0)|| + \\ & ||\mathbf{A}\mathbf{M}^{-1}||\, ||\mathbf{M}||\, ||\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_0|| + \\ & ||\mathbf{A}\mathbf{M}^{-1}||\, ||\, |\hat{\mathbf{L}}|\, |\hat{\mathbf{D}}|\, |\hat{\mathbf{L}}^T|\, ||\, ||\mathbf{M}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_0)|| \Big\} + \mathcal{O}(\varepsilon^2). \end{aligned}$$

Roundoff error right preconditioned GMRES

As we did for FGMRES, if

$$\boxed{c(n)\varepsilon \, \| \, |\hat{\mathbf{L}}| \, |\hat{\mathbf{D}}| \, |\hat{\mathbf{L}}^T| \, \| < \tau}$$

Roundoff error right preconditioned GMRES

As we did for FGMRES, if

$$\boxed{c(n)\varepsilon \, \| \, |\hat{\mathbf{L}}| \, |\hat{\mathbf{D}}| \, |\hat{\mathbf{L}}^T| \, \| < \tau}$$

we can prove that $\exists k^*$ s.t $\forall k \geq k^*$ the right preconditioned GMRES computes a $\bar{\mathbf{x}}_k$ s.t.

$$\begin{aligned} \|\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_k\| \leq \quad & c(n,k)\,\varepsilon \Big[\, \|\mathbf{b}\| + \|\mathbf{A}\| \, \|\bar{\mathbf{x}}_0\| + \\ & \|\mathbf{A}\| \, \|\bar{\mathbf{Z}}_k\| \, \| \, \mathbf{M}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_0)\| + \\ & \| \, |\hat{\mathbf{L}}| \, |\hat{\mathbf{D}}| \, |\hat{\mathbf{L}}^T| \, \| \, \|\mathbf{M}\,(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_0)\| \, \Big] + \mathcal{O}(\varepsilon^2). \end{aligned}$$

Mixed precision arithmetic

- Very fast 32-bit arithmetic unit
  $\mathbf{M}$ is the $fl(\mathbf{LU})$ of $\mathbf{A}$ and $||\mathbf{M} - \mathbf{A}|| \leq c(N)\sqrt{\varepsilon}\,||\mathbf{A}||$
  $(\varepsilon = 2.2 \times 10^{-16})$

Mixed precision arithmetic

- Very fast 32-bit arithmetic unit
  $\mathbf{M}$ is the $fl(\mathbf{LU})$ of $\mathbf{A}$ and $||\mathbf{M} - \mathbf{A}|| \leq c(N)\sqrt{\varepsilon}\,||\mathbf{A}||$
  $(\varepsilon \;= 2.2 \times 10^{-16})$
- We use 32-bit arithmetic for factorization and triangular solves

Mixed precision arithmetic

- Very fast 32-bit arithmetic unit
  $\mathbf{M}$ is the $fl(\mathbf{LU})$ of $\mathbf{A}$ and $||\mathbf{M} - \mathbf{A}|| \leq c(N)\sqrt{\varepsilon}||\mathbf{A}||$
  $(\varepsilon = 2.2 \times 10^{-16})$
- We use 32-bit arithmetic for factorization and triangular solves
- If $\kappa(\mathbf{A})\sqrt{\varepsilon} > 1$ then Iterative Refinement may not converge.
  FGMRES does

Mixed precision arithmetic

- Very fast 32-bit arithmetic unit
  $\mathbf{M}$ is the $fl(\mathbf{LU})$ of $\mathbf{A}$ and $||\mathbf{M}-\mathbf{A}|| \leq c(N)\sqrt{\varepsilon}||\mathbf{A}||$
  $(\varepsilon = 2.2 \times 10^{-16})$
- We use 32-bit arithmetic for factorization and triangular solves
- If $\kappa(\mathbf{A})\sqrt{\varepsilon} > 1$ then Iterative Refinement may not converge.
  FGMRES does
- $||\widehat{\mathbf{W}}_k|| \leq \sqrt{\varepsilon}\, c(N)||\mathbf{A}|| < 1$ and
  $||\mathbf{M}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_0)|| \leq ||\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_k|| + \mathcal{O}(\sqrt{\varepsilon}) \Rightarrow$ FGMRES backward stable

Mixed precision arithmetic

- Very fast 32-bit arithmetic unit
  $\mathbf{M}$ is the $fl(\mathbf{LU})$ of $\mathbf{A}$ and $||\mathbf{M}-\mathbf{A}|| \le c(N)\sqrt{\varepsilon}||\mathbf{A}||$
  $(\varepsilon = 2.2 \times 10^{-16})$
- We use 32-bit arithmetic for factorization and triangular solves
- If $\kappa(\mathbf{A})\sqrt{\varepsilon} > 1$ then Iterative Refinement may not converge. FGMRES does
- $||\widehat{\mathbf{W}}_k|| \le \sqrt{\varepsilon}\, c(N)||\mathbf{A}|| < 1$ and $||\mathbf{M}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_0)|| \le ||\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_k|| + \mathcal{O}(\sqrt{\varepsilon}) \Rightarrow$ FGMRES backward stable
- GMRES is not backward stable

Improved error analysis for FGMRES

If we apply Flex-GMRES to solve the system, using finite-precision arithmetic conforming to IEEE standard with relative precision ε and under the following hypotheses:

$$2.12(n+1)\varepsilon < 0.01 \qquad \text{and} \qquad c_0(n)\varepsilon\, \kappa(\mathbf{C}^{(k)}) < 0.1\ \forall k$$

where

$$c_0(n) = 18.53 n^{\frac{3}{2}}$$

and

$$|\bar{s}_k| < 1 - \varepsilon\,,\ \forall k,$$

where $\bar{s}_k$ are the sines computed during the Givens algorithm applied to $\bar{\mathbf{H}}_k$ in order to compute $\bar{\mathbf{y}}_k$, then there exists $\hat{k}$, $\hat{k} \leq n$ such that, $\forall k \geq \hat{k}$, we have

$$\begin{aligned}||\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_k|| \quad &\leq \quad c_1(n,k)\varepsilon \Big(||\mathbf{b}|| + ||\mathbf{A}||\, ||\bar{\mathbf{x}}_0|| + \\ &\qquad ||\mathbf{A}||\, ||\, |\bar{\mathbf{Z}}_k|\, |\bar{\mathbf{y}}_k|\, || + ||\mathbf{A}\bar{\mathbf{Z}}_k||\, ||\bar{\mathbf{y}}_k|| \Big) + \mathcal{O}(\varepsilon^2).\end{aligned}$$

Test Problems

	n	nnz	Description
CONT_201	80595	239596	KKT matrix Convex QP (M2)
CONT_300	180895	562496	KKT matrix Convex QP (M2)
TUMA_1	22967	76199	Mixed-Hybrid finite-element

Test problems

MA57 tests

	n	nnz(L)+nnz(D)	Factorization time
CONT_201	80595	9106766	9.0 sec
CONT_300	180895	22535492	28.8 sec

MA57 without static pivot

MA57 tests

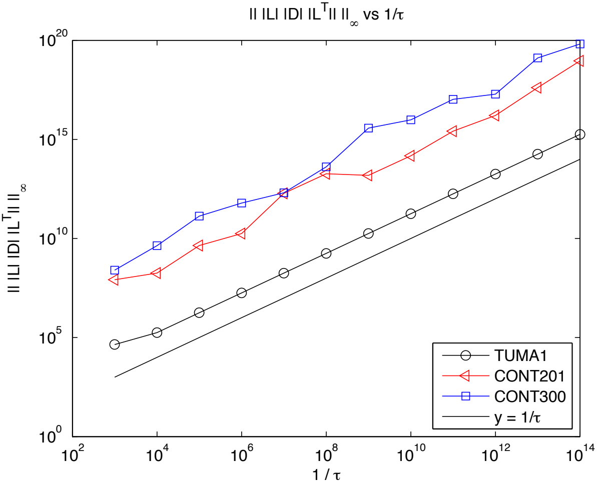
	n	nnz(L)+nnz(D)	Factorization time
CONT_201	80595	9106766	9.0 sec
CONT_300	180895	22535492	28.8 sec

MA57 without static pivot

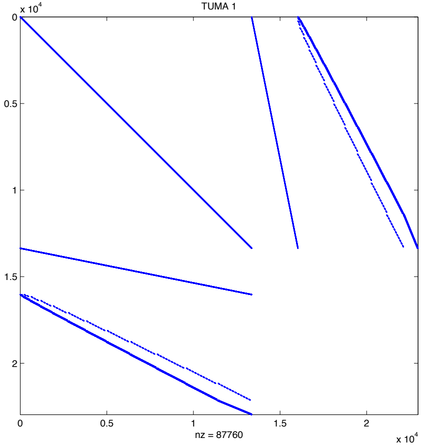
	nnz(L)+nnz(D)+ FGMRES (#it)	Factorization time	# static pivots
CONT_201	5563735 (6)	3.1 sec	27867
CONT_300	12752337 (8)	8.9 sec	60585

MA57 with static pivot $\tau = 10^{-8}$

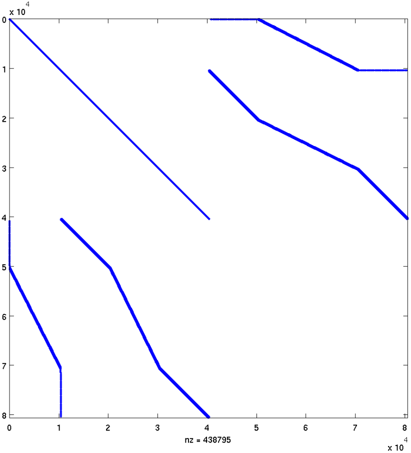
$\| \, |\hat{\mathbf{L}}| \, |\hat{\mathbf{D}}| \, |\hat{\mathbf{L}}^T| \, \|$ vs $1/\tau$



Test Problems: TUMA 1



Test Problems: CONT-201



Numerical experiments: TUMA 1

	$\frac{\|\|b - A\bar{\mathbf{x}}_k\|\|}{\|\|b\|\| + \|\|A\|\|\|\|\bar{\mathbf{x}}_k\|\|}$				$\|\|M(\bar{\mathbf{x}}_k - \bar{x}_0)\|\|$		
τ	IR	GMRES	FGMRES	$\|\|Z_k\|\|$	GMRES	FGMRES	$\|\| \|L\| \|D\| \|L^T\| \|\|$
1.0e-03	3.0e-03	1.0e-14	7.2e-17	1.2e+02	3.5e-03	3.5e-03	4.4e+04
1.0e-04	5.3e-17	1.8e-16	3.1e-17	4.7e+01	4.4e-04	4.4e-04	1.8e+05
1.0e-05	5.1e-17	1.3e-16	1.9e-17	4.4e+01	4.5e-05	4.5e-05	1.8e+06
1.0e-06	1.5e-16	1.3e-16	1.9e-17	4.4e+01	4.5e-06	4.5e-06	1.8e+07
1.0e-07	1.8e-17	1.2e-16	2.0e-17	4.3e+01	4.5e-07	4.5e-07	1.8e+08
1.0e-08	1.7e-17	1.3e-16	1.8e-17	4.3e+01	4.5e-08	4.5e-08	1.8e+09
1.0e-09	1.8e-17	2.8e-15	1.8e-17	2.6e+01	4.0e-08	4.0e-08	1.8e+10
1.0e-10	1.7e-17	4.2e-13	1.8e-17	8.8e+00	4.0e-07	4.0e-07	1.8e+11
1.0e-11	6.7e-17	1.0e-10	6.2e-17	6.8e+00	4.0e-06	4.0e-06	1.8e+12
1.0e-12	2.1e-17	1.0e-08	2.2e-17	3.2e+01	4.3e-05	4.3e-05	1.8e+13
1.0e-13	2.0e-17	2.4e-07	1.9e-17	1.3e+02	3.9e-04	3.9e-04	1.8e+14
1.0e-14	8.6e-17	8.6e-06	2.1e-17	1.8e+02	4.3e-03	4.3e-03	1.8e+15

TUMA 1 results

Numerical experiments: CONT_201

	$\frac{\|\|b - A\bar{\mathbf{x}}_k\|\|}{\|\|b\|\| + \|\|A\|\|\|\|\bar{\mathbf{x}}_k\|\|}$				$\|\|M(\bar{\mathbf{x}}_k - \bar{x}_0)\|\|$		
τ	IR	GMRES	FGMRES	$\|\|Z_k\|\|$	GMRES	FGMRES	$\|\| \|L\| \|D\| \|L^T\| \|\|$
1.0e-03	4.0e-04	1.8e-05	9.8e-06	*	7.1e-04	1.5e-04	8.3e+07
1.0e-04	4.0e-05	2.0e-07	2.0e-07	*	1.5e-05	1.9e-05	1.8e+08
1.0e-05	3.5e-06	1.8e-12	1.1e-16	4.1e+05	5.9e-06	1.3e-05	4.4e+09
1.0e-06	3.5e-07	1.1e-11	2.1e-16	2.7e+06	7.8e-07	7.8e-07	1.8e+10
1.0e-07	4.0e-08	4.8e-11	1.8e-16	1.4e+08	8.7e-08	8.7e-08	1.9e+12
1.0e-08	3.8e-13	2.7e-10	5.8e-17	2.1e+07	1.3e-06	1.3e-06	1.8e+13
1.0e-09	5.5e-17	1.8e-09	4.5e-17	1.1e+07	1.3e-06	1.3e-06	1.5e+13
1.0e-10	7.7e-17	3.2e-09	7.2e-17	3.4e+05	9.2e-06	9.2e-06	1.5e+14
1.0e-11	4.6e-17	2.1e-09	4.5e-17	1.9e+03	2.8e-04	2.8e-04	2.6e+15
1.0e-12	5.2e-17	4.5e-07	3.8e-17	2.0e+02	9.5e-04	9.5e-04	1.6e+16
1.0e-13	1.3e-16	1.3e-04	2.6e-16	1.6e+02	1.1e-02	1.1e-02	4.1e+17
1.0e-14	1.2e-03	2.3e-01	2.5e-14	4.3e+02	1.9e-02	1.0e-02	9.2e+18

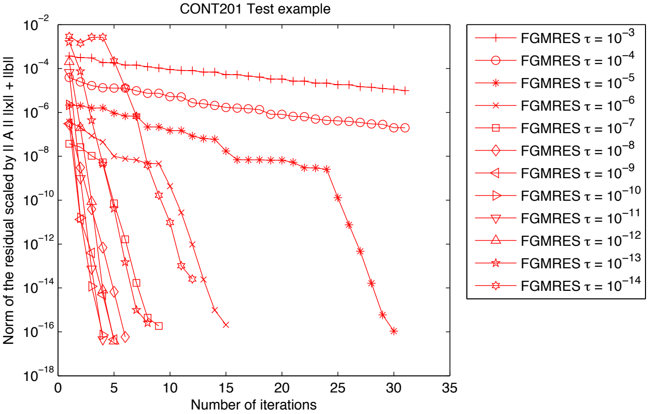
CONT_201 results

Numerical experiments: CONT_300

	$\frac{\lVert b - A\bar{\mathbf{x}}_k \rVert}{\lVert b \rVert + \lVert A \rVert \lVert \bar{\mathbf{x}}_k \rVert}$				$\lVert M(\bar{\mathbf{x}}_k - \bar{x}_0) \rVert$		
τ	IR	GMRES	FGMRES	$\lVert Z_k \rVert$	GMRES	FGMRES	$\lVert \, \lvert L \rvert \, \lvert D \rvert \, \lvert L^T \rvert \, \rVert$
1.0e-03	3.8e-04	3.6e-05	2.5e-05	*	8.7e-04	1.3e-04	2.5e+08
1.0e-04	3.6e-05	5.5e-07	5.5e-07	*	6.5e-05	2.8e-05	4.3e+09
1.0e-05	4.3e-06	8.7e-09	8.7e-09	*	3.7e-06	6.1e-06	1.4e+11
1.0e-06	3.7e-07	6.9e-11	1.4e-16	3.0e+06	5.7e-07	9.8e-07	6.2e+11
1.0e-07	6.8e-08	2.1e-10	8.2e-17	7.6e+06	2.3e-07	2.3e-07	2.0e+12
1.0e-08	2.1e-09	1.4e-08	1.2e-16	7.5e+07	1.8e-06	1.8e-06	4.1e+13
1.0e-09	1.1e-16	1.6e-05	8.8e-17	3.7e+07	2.8e-04	2.8e-04	3.7e+15
1.0e-10	3.9e-17	6.8e-07	4.1e-17	3.8e+05	3.6e-04	3.6e-04	9.6e+15
1.0e-11	4.0e-17	1.6e-06	8.7e-17	1.4e+03	5.3e-03	5.3e-03	1.0e+17
1.0e-12	7.3e-17	1.1e-06	2.7e-16	1.5e+02	1.0e-02	1.0e-02	1.9e+17
1.0e-13	1.8e-16	3.4e-03	9.2e-16	1.3e+02	1.9e-01	1.9e-01	1.3e+19
1.0e-14	1.1e-15	1.4e-01	1.8e-14	2.1e+02	4.7e-02	4.7e-02	6.6e+19

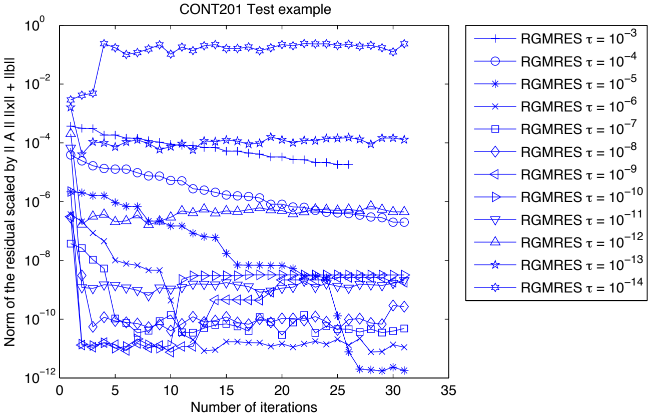
CONT_300 results

Numerical experiments



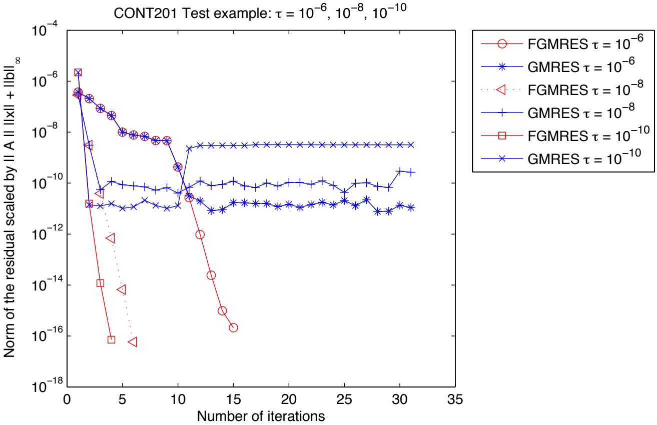
FGMRES on CONT-201 test example

Numerical experiments



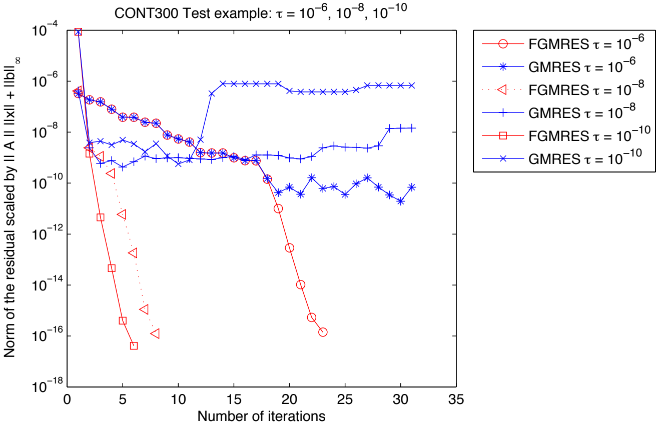
GMRES on CONT-201 test example

Numerical experiments



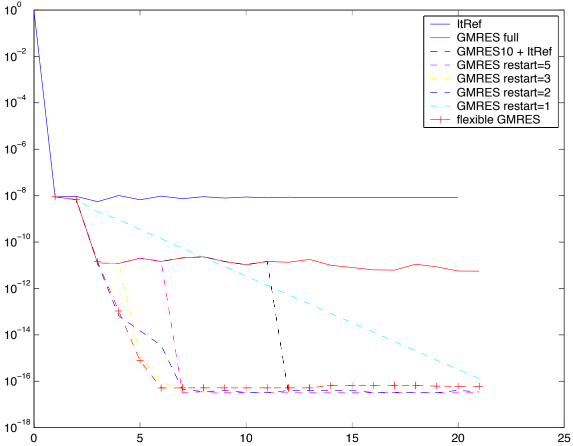
GMRES vs. FGMRES on CONT-201 test example:
$\tau = 10^{-6}, 10^{-8}, 10^{-10}$

Numerical experiments



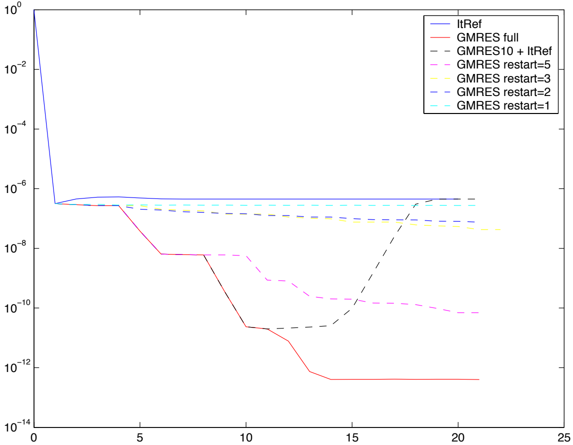
GMRES vs. FGMRES on CONT-300 test example:
$\tau = 10^{-6}, 10^{-8}, 10^{-10}$

Numerical experiments



Restarted GMRES vs. FGMRES on CONT-201 test example: $\tau = 10^{-8}$

Numerical experiments



Restarted GMRES on CONT-201 test example: $\tau = 10^{-6}$

IR vs FGMRES

	Iterative refinement		FGMRES			
Matrix	Total it	RR	Total / inner	RR	$\|\mathbf{A}\bar{\mathbf{Z}}_{\hat{k}}\|$	$\|\,\|\bar{\mathbf{Z}}_{\hat{k}}\|\,\|\bar{\mathbf{y}}_{\hat{k}}\|\,\|$
bcsstk20	30	2.1e-15	2 / 2	1.4e-11	1.7e+00	4.6e+02
$n = 485$			4 / 2	3.4e-14	1.6e+00	3.8e-01
$\kappa(A) \approx 4 \times 10^{12}$			6 / 2	7.2e-17	1.6e+00	5.6e-04
bcsstm27	22	1.6e-15	2 / 2	5.8e-11	1.7e+00	2.7e+01
$n = 1224$			4 / 2	1.8e-11	6.3e-01	1.3e+00
$\kappa(A) \approx 5 \times 10^{9}$			6 / 2	6.0e-13	2.0e+00	7.6e-02
			8 / 2	1.5e-13	1.7e+00	1.0e-02
			10 / 2	1.2e-14	1.7e+00	1.9e-03
			12 / 2	2.6e-15	1.8e+00	1.7e-04
			14 / 2	1.8e-16	1.6e+00	4.3e-05
s3rmq4m1	16	2.2e-15	2 / 2	3.5e-11	1.0e+00	8.6e+01
$n = 5489$			4 / 2	2.1e-13	1.1e+00	3.2e-01
$\kappa(A) \approx 4 \times 10^{9}$			6 / 2	4.5e-15	1.7e+00	6.4e-03
			8 / 2	1.1e-16	1.6e+00	1.3e-04
s3dkq4m2 $n = 90449$ $\kappa(\mathbf{A}) \approx 7 \times 10^{10}$	53	1.1e-10	10 / 10	6.3e-17	1.2e+00	1.2e+03

Summary

- IR with static pivoting is very sensitive to τ and not robust

Summary

- IR with static pivoting is very sensitive to τ and not robust
- GMRES is also sensitive and not robust

Summary

- IR with static pivoting is very sensitive to τ and not robust
- GMRES is also sensitive and not robust
- FGMRES is robust and less sensitive (see roundoff analysis)

Summary

- IR with static pivoting is very sensitive to τ and not robust
- GMRES is also sensitive and not robust
- FGMRES is robust and less sensitive (see roundoff analysis)
- Gains from restarting. Makes GMRES more robust, saves storage in FGMRES (but not really needed)

Summary

- IR with static pivoting is very sensitive to τ and not robust
- GMRES is also sensitive and not robust
- FGMRES is robust and less sensitive (see roundoff analysis)
- Gains from restarting. Makes GMRES more robust, saves storage in FGMRES (but not really needed)
- Understanding of why $\tau \approx \sqrt{\varepsilon}$ is best.