Science & Technology Facilities Council
Rutherford Appleton Laboratory

Generalized Goulb-Kahan bidiagonalization and stopping criteria

Mario Arioli

m.arioli@rl.ac.uk

STFC-Rutherford Appleton Laboratory

Outline

- Norms and duality
- The minimization problem
- G-K bidiagonalization
- Stopping criteria
- inf-sup conditions and Generalized singular values
- Numerical examples
- Summary and open problems

Linear operators

Let $\mathbf{M} \in \mathbf{R}^{m \times m}$ and $\mathbf{N} \in \mathbf{R}^{n \times n}$ be symmetric positive definite matrices, and let $\mathbf{A} \in \mathbf{R}^{m \times n}$ be a full rank matrix.

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

$$\mathcal{M}' = \{\mathbf{w} \in \mathbf{R}^m; \|\mathbf{w}\|^2_{\mathbf{M}^{-1}} = \mathbf{w}^T\mathbf{M}^{-1}\mathbf{w}\},\ \mathcal{N}' = \{\mathbf{y} \in \mathbf{R}^n; \|\mathbf{y}\|^2_{\mathbf{N}^{-1}} = \mathbf{y}^T\mathbf{N}^{-1}\mathbf{y}\}$$

Linear operators

Let $\mathbf{M} \in \mathbf{R}^{m \times m}$ and $\mathbf{N} \in \mathbf{R}^{n \times n}$ be symmetric positive definite matrices, and let $\mathbf{A} \in \mathbf{R}^{m \times n}$ be a full rank matrix.

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

$$\mathcal{M}' = \{\mathbf{w} \in \mathbf{R}^m; \|\mathbf{w}\|^2_{\mathbf{M}^{-1}} = \mathbf{w}^T\mathbf{M}^{-1}\mathbf{w}\},\ \mathcal{N}' = \{\mathbf{y} \in \mathbf{R}^n; \|\mathbf{y}\|^2_{\mathbf{N}^{-1}} = \mathbf{y}^T\mathbf{N}^{-1}\mathbf{y}\}$$

$$\langle \mathbf{v}, \mathbf{A}\mathbf{q} \rangle_{\mathcal{M},\mathcal{M}'} = \mathbf{v}^T\mathbf{A}\mathbf{q}, \quad \mathbf{A}\mathbf{q} \in \mathcal{L}(\mathcal{M})\ \forall \mathbf{q} \in \mathcal{N}.$$

Linear operators

Let $\mathbf{M} \in \mathbf{R}^{m \times m}$ and $\mathbf{N} \in \mathbf{R}^{n \times n}$ be symmetric positive definite matrices, and let $\mathbf{A} \in \mathbf{R}^{m \times n}$ be a full rank matrix.

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

$$\mathcal{M}' = \{\mathbf{w} \in \mathbf{R}^m; \|\mathbf{w}\|^2_{\mathbf{M}^{-1}} = \mathbf{w}^T\mathbf{M}^{-1}\mathbf{w}\},\ \mathcal{N}' = \{\mathbf{y} \in \mathbf{R}^n; \|\mathbf{y}\|^2_{\mathbf{N}^{-1}} = \mathbf{y}^T\mathbf{N}^{-1}\mathbf{y}\}$$

$$\langle \mathbf{v}, \mathbf{A}\mathbf{q} \rangle_{\mathcal{M},\mathcal{M}'} = \mathbf{v}^T\mathbf{A}\mathbf{q}, \quad \mathbf{A}\mathbf{q} \in \mathcal{L}(\mathcal{M})\ \forall \mathbf{q} \in \mathcal{N}.$$

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

$$\langle \mathbf{A}^\bigstar\mathbf{g}, \mathbf{f} \rangle_{\mathcal{N}',\mathcal{N}} = \mathbf{f}^T\mathbf{A}^T\mathbf{g}, \quad \mathbf{A}^T\mathbf{g} \in \mathcal{L}(\mathcal{N})\ \forall \mathbf{g} \in \mathcal{M},$$

Generalized SVD

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

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

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

Generalized SVD

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

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

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

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

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

Generalized SVD

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

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

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

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

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

The generalized singular values are the standard singular values of

$$\tilde{\mathbf{A}} = \mathbf{M}^{-1/2} \mathbf{A} \mathbf{N}^{-1/2}.$$

The generalized 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 $\tilde{\mathbf{A}}$.

Remark on inf-sup

We point out that the necessary and sufficient conditions, based on the *inf-sup* condition (Brezzi, 1974 and 2002, Brezzi-Fortin 1991), that guarantee both existence and unicity of the solution and the stability, are equivalent to impose that the generalized singular values σ_i of $\mathbf{A}$ are in the interval (a, b) with $0 < a < b$ and a and b independent of the dimensions n and m. This also implies that the generalized condition number $\kappa(\mathbf{A}) = \dfrac{\sigma_1}{\sigma_n}$ is independent of n and m.

Problem

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

where $\mathbf{M}$ is a nonsingular symmetric and positive definite matrix.
Several problems can be reduced to this case.

Problem

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

where $\mathbf{M}$ is a nonsingular symmetric and positive definite matrix. Several problems can be reduced to this case. The general problem

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

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

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

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

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

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

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

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

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 2

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

Algorithm 2

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

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

Algorithm 2

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

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

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 $\mathbf{U}$ by

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

u

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

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

$$\mathbf{D}_k = \mathbf{B}_k^{-T} \mathbf{Q}_k^T$$

$$\mathbf{d}_1 = \frac{\mathbf{q}_1}{\alpha_1}$$

$$\mathbf{d}_{i+1} = \frac{\mathbf{q}_{i+1} - \beta_{i+1}\mathbf{d}_i}{\alpha_{i+1}} \qquad i = 1, \ldots, n,$$

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

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

Craig's variant algorithm

procedure $[\mathbf{U}, \mathbf{V}, \mathbf{B}, \mathbf{u}, \mathbf{p}]$ = G-K_bidiagonalization$(\mathbf{A}, \mathbf{M}, \mathbf{N}, \mathbf{b}, maxit)$;

$\beta_0 = \|\mathbf{b}\|_{\mathbf{N}}$; $\mathbf{q}_1 = \mathbf{N}^{-1}\mathbf{b}/\beta_0$;

$\mathbf{w} = \mathbf{M}^{-1}\mathbf{A}\mathbf{q}_1$; $\alpha_1 = \mathbf{w}^T\mathbf{M}\mathbf{w}$; $\mathbf{v}_1 = \mathbf{w}/\sqrt{\alpha_1}$;

$\zeta_1 = -\beta_0/\alpha_1$; $\mathbf{d}_1 = \mathbf{q}_1/\alpha_1$; $\mathbf{p}^{(1)} = -\zeta_1\mathbf{d}_1$

k = 0; it = 0; convergence = false;

while convergence = false and $it < maxit$

$k = k + 1$; $it = it + 1$;

$\mathbf{g} = \mathbf{N}^{-1}\left(\mathbf{A}^T\mathbf{v}_k - \alpha_i\mathbf{N}\mathbf{q}_k\right)$; $\beta_{k+1} = \mathbf{g}^T\mathbf{N}\mathbf{g}$;

$\mathbf{q}_{k+1} = \mathbf{g}\,\sqrt{\beta_{k+1}}$;

$\mathbf{w} = \mathbf{M}^{-1}\left(\mathbf{A}\mathbf{q}_{k+1} - \beta_{k+1}\mathbf{M}\mathbf{v}_k\right)$; $\alpha_{k+1} = \mathbf{w}^T\mathbf{M}\mathbf{w}$;

$\mathbf{v}_{k+1} = \mathbf{w}/\sqrt{\alpha_{k+1}}$; $\zeta_{k+1} = \dfrac{\beta_k}{\alpha_{k+1}}\zeta_k$;

$\mathbf{d}_{k+1} = \left(\mathbf{q}_{k+1} - \beta_{k+1}\mathbf{d}_k\right)/\alpha_{k+1}$;

$\mathbf{u}^{(k+1} = \mathbf{u}^{(k)} + \zeta_{k+1}\mathbf{v}_{k+1}$; $\mathbf{p}^{(k+1} = \mathbf{p}^{(k)} - \zeta_{k+1}\mathbf{d}_{k+1}$;

[convergence] = check$(\mathbf{z}_k, \ldots)$

end while;

end procedure.

Stopping criteria

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

Stopping criteria

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

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

Stopping criteria

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

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

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

Lower bound

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

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

Lower bound

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

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

The procedure "*check*$(\mathbf{z}_k, \ldots)$" can then specialized as

procedure [convergence] = check$(\mathbf{z}_k, k, d, \tau)$
    convergence = false;
    if $k > d$ then
        $\xi^2 = \sum_{j=k-d+1}^{k} \zeta_j^2$;
        if $\xi \leq \tau$ then;
            convergence = true;
        end if;
    end if;
end procedure.

Lower bound

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

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

The procedure "*check*$(\mathbf{z}_k, \dots)$" can then specialized as

procedure [convergence] = check$(\mathbf{z}_k, k, d, \tau)$

```
    convergence = false;
    if k > d then
        ξ² = Σ_{j=k-d+1}^{k} ζ_j²;
        if ξ ≤ τ then;
            convergence = true;
        end if;
    end if;
end procedure.
```

$$\|\mathbf{e}^{(k)}\|_{\mathbf{M}}^2 = \sum_{j=k+1}^{n} \zeta_j^2 = \|b\|_{\mathbf{N}}^2 \left[\left(\mathbf{T}^{-1}\right)_{1,1} - \left(\mathbf{T}_k^{-1}\right)_{1,1} \right], \quad (\mathbf{T} = \mathbf{B}^T\mathbf{B})$$

(Golub and Meurant 2010)

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

$$\hat{\mathbf{T}}_{k+1} = \begin{bmatrix} \mathbf{T}_k & \alpha_k \beta_k \mathbf{e}_k \\ \alpha_k \beta_k \mathbf{e}_k^T & \omega_{k+1} \end{bmatrix},$$

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

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

We can recursively compute $\delta(a^2)_k$ and ω_{k+1} by using the Cholesky decomposition.

Upper bound

We obtain the following realization of the procedure "*check*$(\mathbf{z}_k, \ldots)$"

procedure [convergence] = checkGR$(\mathbf{z}_k, k, d, \tau, a, \|b\|_{\mathbf{N}}, \mathbf{B}_k)$

convergence = false;

if $k = 1$ then

$\bar{d}_1 = \alpha_1^2 + \beta_1^2 - a^2$;

else

$\bar{d}_k = \alpha_k^2 + \beta_k^2 - \varpi_{k-1}$;

end if;

$$\varpi_k = a^2 + \frac{\alpha_k^2 \beta_k^2}{\bar{d}_k}; \quad \varphi_k = \frac{\beta_k^2 \zeta_k^2}{\sqrt{\bar{d}_k + a^2 - \beta_k^2}};$$

if $k > d$ then

$\xi^2 = \sum_{j=k-d+1}^{k} \zeta_j^2$; $\quad \Xi^2 = \xi^2 + \varphi_k$;

if $\Xi \leq \tau$ then;

convergence = true;

end if;

end if;

end procedure.

Test problems

The Poisson problem with mixed boundary conditions on $\Omega = (0,1) \times (0,1)$:

$$
\begin{aligned}
-\nabla \cdot \nabla u &= f \quad \text{in } \Omega, \\
\frac{\partial u}{\partial \mathbf{n}} &= 0 \quad \text{on } \partial_N \Omega = \{0 \times (0,1)\} \cup \{1 \times (0,1)\}, \\
u &= 0 \quad \text{on } \partial_D \Omega = \{(0,1) \times 0\} \\
u &= 1 \quad \text{on } \partial_D \Omega = \{(0,1) \times 1\}.
\end{aligned}
$$

where $\mathbf{n}$ is the external normal to the domain.

Test problems

- The Poisson problem with mixed boundary conditions on $\Omega = (0,1) \times (0,1)$:

$$\begin{aligned} -\nabla \cdot \nabla u &= f \quad \text{in } \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} &= 0 \quad \text{on } \partial_N \Omega = \{0 \times (0,1)\} \cup \{1 \times (0,1)\}, \\ u &= 0 \quad \text{on } \partial_D \Omega = \{(0,1) \times 0\} \\ u &= 1 \quad \text{on } \partial_D \Omega = \{(0,1) \times 1\}. \end{aligned}$$

  where $\mathbf{n}$ is the external normal to the domain.

- The Poisson equation with Neumann zero boundary conditions on a domain $\Omega = (0,1) \times (0,1)$:

$$\begin{aligned} -\nabla \cdot \nabla u &= f \quad \text{in } \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

  where $\mathbf{n}$ is the external normal to the domain and f has zero mean.

Test problems

- The Poisson problem with mixed boundary conditions on $\Omega = (0,1) \times (0,1)$:

$$\begin{aligned} -\nabla \cdot \nabla u &= f \quad \text{in } \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} &= 0 \quad \text{on } \partial_N \Omega = \{0 \times (0,1)\} \cup \{1 \times (0,1)\}, \\ u &= 0 \quad \text{on } \partial_D \Omega = \{(0,1) \times 0\} \\ u &= 1 \quad \text{on } \partial_D \Omega = \{(0,1) \times 1\}. \end{aligned}$$

  where $\mathbf{n}$ is the external normal to the domain.

- The Poisson equation with Neumann zero boundary conditions on a domain $\Omega = (0,1) \times (0,1)$:

$$\begin{aligned} -\nabla \cdot \nabla u &= f \quad \text{in } \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

  where $\mathbf{n}$ is the external normal to the domain and f has zero mean.

- The Stokes problem on a domain with a step: Ω is the L-shaped region generated by taking the complement in $(1, L) \times (1, 1)$ of the quadrant $(1, 0] \times (1, 0]$.

Test problems

- The Poisson problem with mixed boundary conditions on $\Omega = (0,1) \times (0,1)$:

$$\begin{aligned} -\nabla \cdot \nabla u &= f \quad \text{in } \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} &= 0 \quad \text{on } \partial_N \Omega = \{0 \times (0,1)\} \cup \{1 \times (0,1)\}, \\ u &= 0 \quad \text{on } \partial_D \Omega = \{(0,1) \times 0\} \\ u &= 1 \quad \text{on } \partial_D \Omega = \{(0,1) \times 1\}. \end{aligned}$$

  where $\mathbf{n}$ is the external normal to the domain.

- The Poisson equation with Neumann zero boundary conditions on a domain $\Omega = (0,1) \times (0,1)$:

$$\begin{aligned} -\nabla \cdot \nabla u &= f \quad \text{in } \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

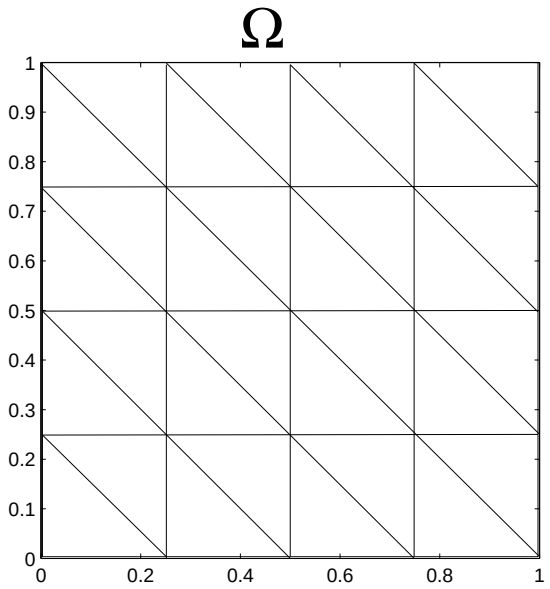
  where $\mathbf{n}$ is the external normal to the domain and f has zero mean.

- The Stokes problem on a domain with a step: Ω is the L-shaped region generated by taking the complement in $(1, L) \times (1, 1)$ of the quadrant $(1, 0] \times (1, 0]$.

- A set of Darcy's problems supplied by the Dept. of Mathematical Modelling in DIAMO, s.e., Straz pod Ralskem, Czech Republic

Ω and mesh for Poisson problems

$$\Omega = (0,1) \times (0,1)$$



An example of uniform triangulation

Poisson with mixed b.c. Problems

The Poisson problem is casted in its dual form as a Darcy's problem:

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

We approximated the spaces $\mathcal{H}$ 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 $\mathcal{H}_h$ of $\mathcal{H}$ is

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

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

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

Poisson with mixed b.c. Data

$h = 2^{-k}$	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 $u(x, y) = x$.

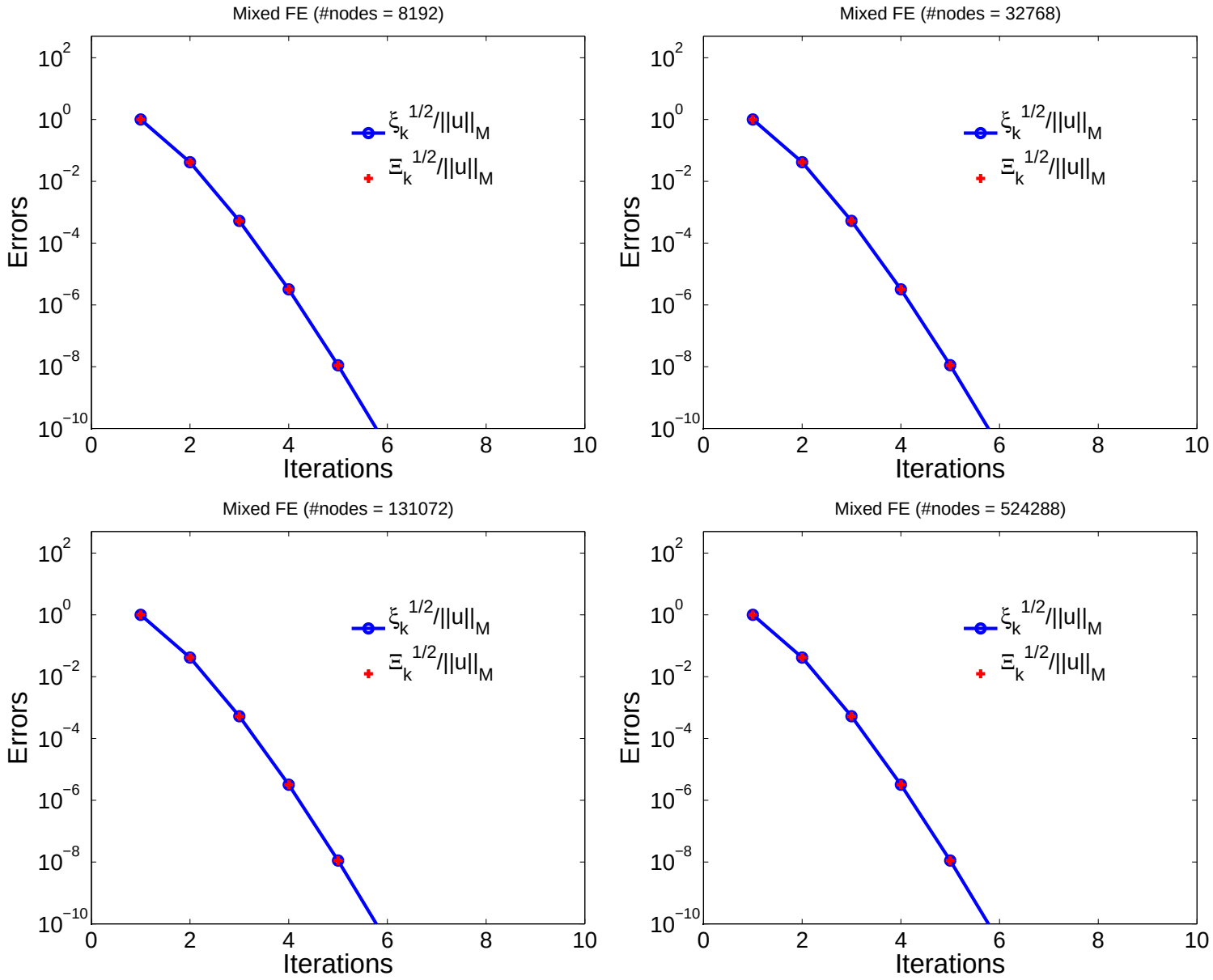
We point out that the pattern of $\mathbf{W}$ is structurally equal to the pattern $\mathbf{A}\mathbf{N}^{-1}\mathbf{A}^T$.

Poisson with mixed b.c. Problems results

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

MIX problems ($d = 5$, $\tau = 10^{-8}$)



Poisson with Neumann b.c. Problem

$$\left.\begin{array}{rcl} \vec{w}(x) + \nabla u & = & 0 \\ \nabla \cdot \vec{w}(x) & = & f \end{array}\right\}.$$

We partition the domain by $\sqrt{n} \times \sqrt{n}$ uniform mesh, where $n = 4^k$ for a fixed k and approximate the derivative by finite differences. The Neumann boundary conditions imply that $\mathbf{w} = 0$ outside Ω.
We point out that the matrix $\mathbf{A}$ is not full rank. We chose the following values for k

$$k = \{5, 6, 7, 8, 9\}\,.$$

In all the 5 cases, the right hand side $\mathbf{b}$ has been chosen with entries

$$b_i = \begin{cases} -1 & i \leq \frac{n}{2}, \\ 1 & i > \frac{n}{2}. \end{cases}$$

Poisson with Neumann b.c. Problem

name	m	n	nnz($\mathbf{M}$)	nnz($\mathbf{E}$)
NFD1	1984	1024	7748	3968
NFD2	8064	4096	31876	16128
NFD3	32512	16384	129284	65024
NFD4	130560	65536	520708	261120
NFD5	523264	262144	2089988	1046528

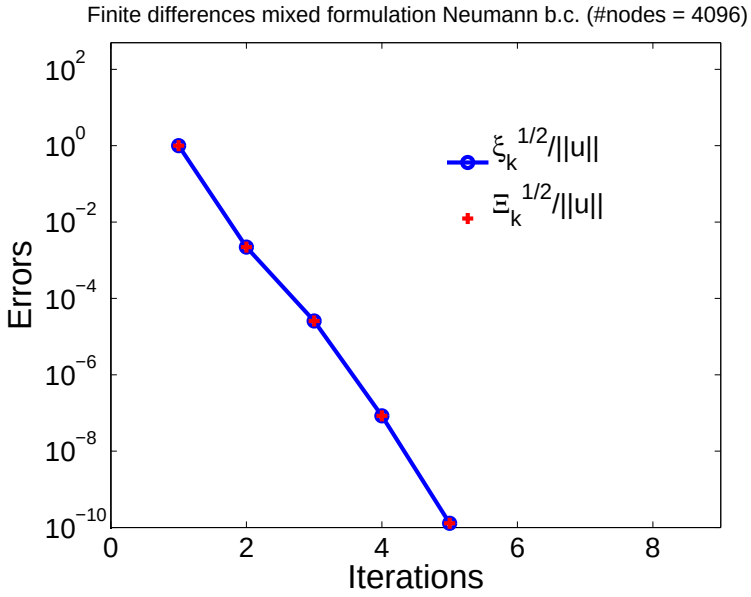
(nnz($\mathbf{M}$) is only for the symmetric part)

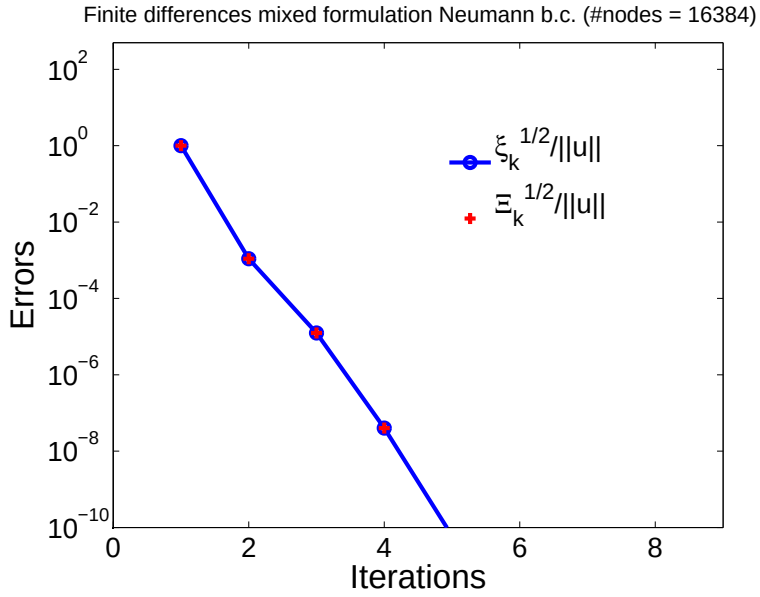
Poisson with Neumann b.c. Problems results

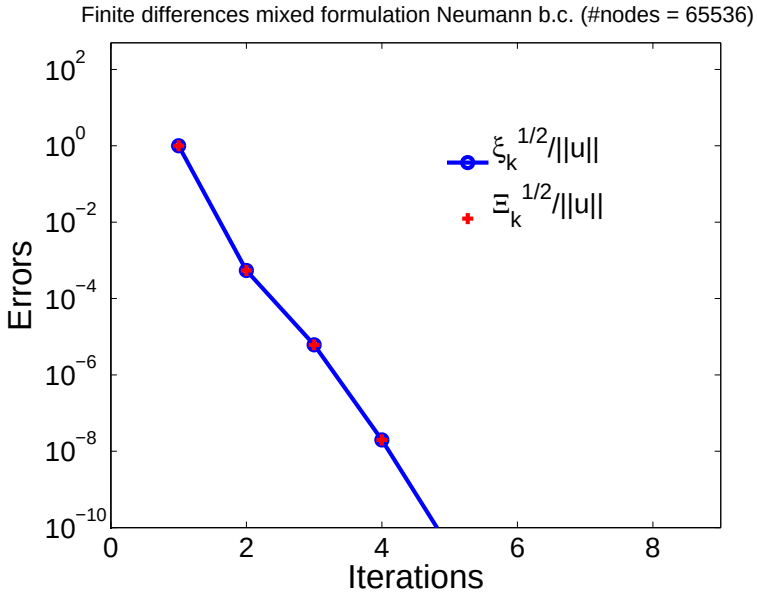
name	# Iter.s	$\|\mathbf{e}^{(k)}\|_2$	$\|\mathbf{A}^T\mathbf{u}^{(k)} - \mathbf{b}\|_2$	$\kappa(\mathbf{B})$
NFD1	9	1.5e-12	3.3e-12	5.5e+04
NFD2	9	1.2e-12	3.1e-13	8.2e+03
NFD3	9	4.7e-12	2.5e-12	4.4e+04
NFD4	9	2.0e-11	2.2e-12	1.7e+04
NFD5	9	9.0e-11	1.2e-13	6.0e+03

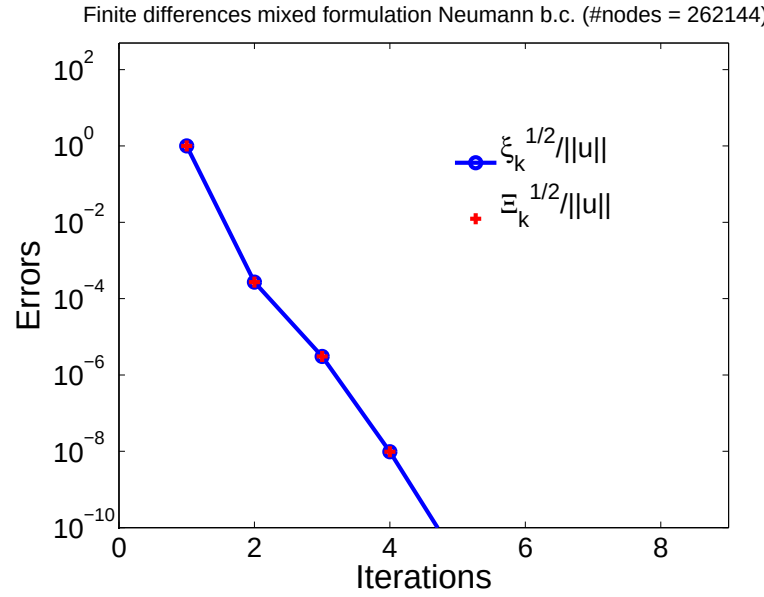
Poisson with Neumann b.c. problems: results ($d = 5$, $\tau = 10^{-8}$).

Neumann problems ($d = 5, \tau = 10^{-8}$)









Stokes Problems

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(M)	nnz(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

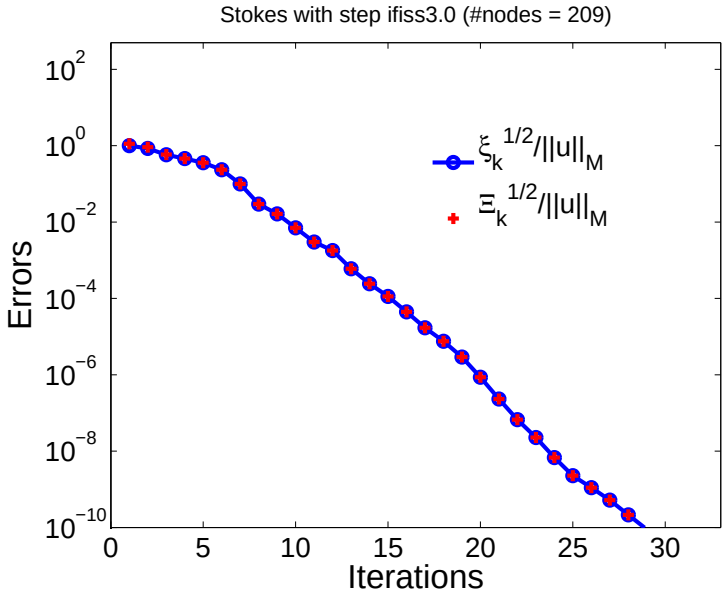
(nnz(M) is only for the symmetric part)

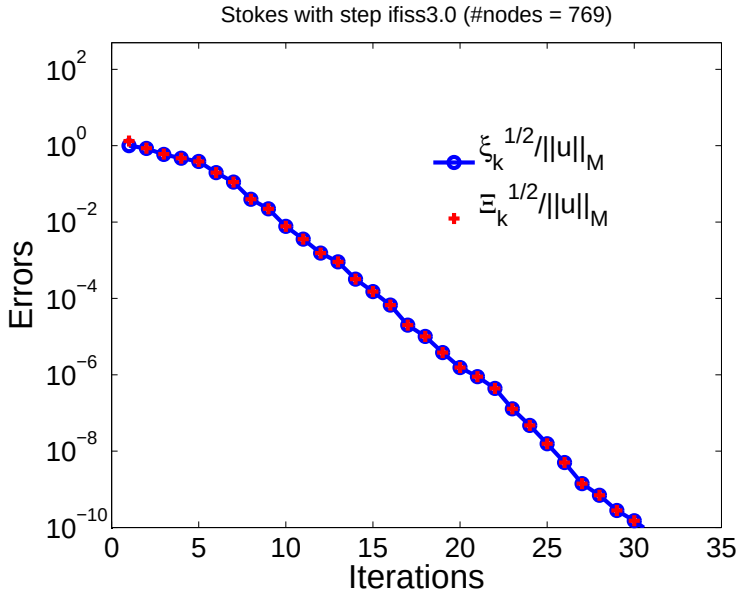
Stokes (Step) Problems results

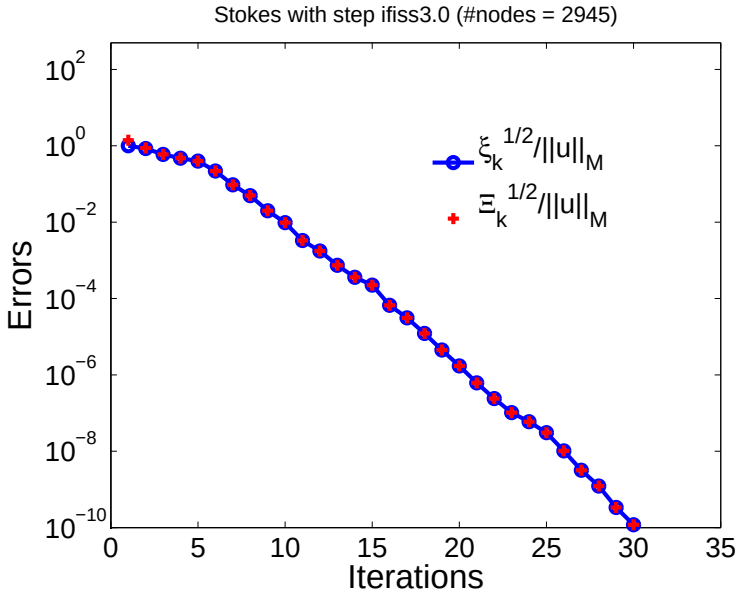
name	# Iter.s	$\|\mathbf{e}^{(k)}\|_2$	$\|\mathbf{A}^T\mathbf{u}^{(k)} - \mathbf{b}\|_2$	$\|\mathbf{p} - \mathbf{p}^{(k)}\|_2$	$\kappa(\mathbf{B})$
Step1	30	6.8e-16	5.1e-16	1.1e-13	7.6
Step2	32	5.4e-14	5.4e-14	5.0e-12	7.7
Step3	34	3.8e-14	2.7e-14	1.0e-11	7.8
Step4	34	5.0e-13	1.3e-13	1.4e-10	7.8
Step5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

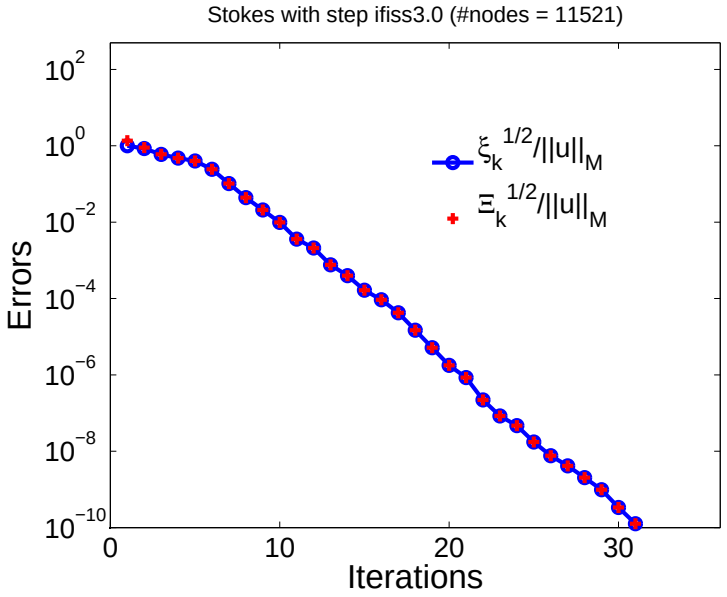
Stokes (Step) problems results ($d = 5$, $\tau = 10^{-8}$).

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







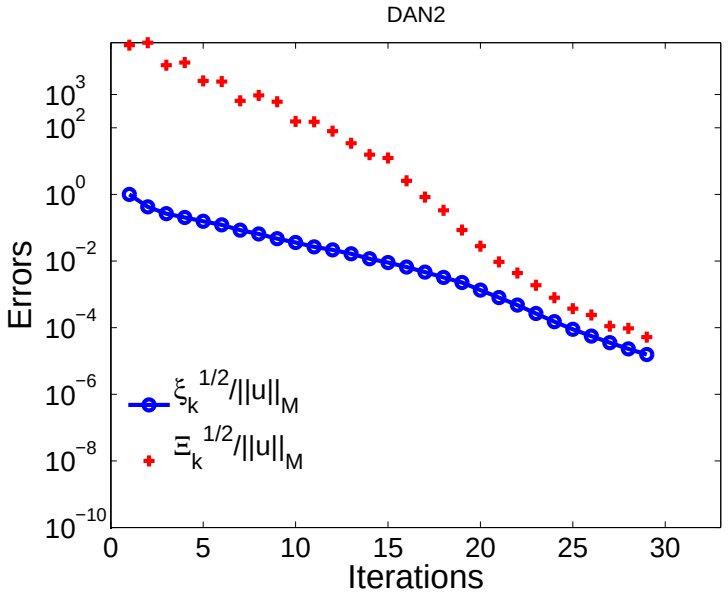


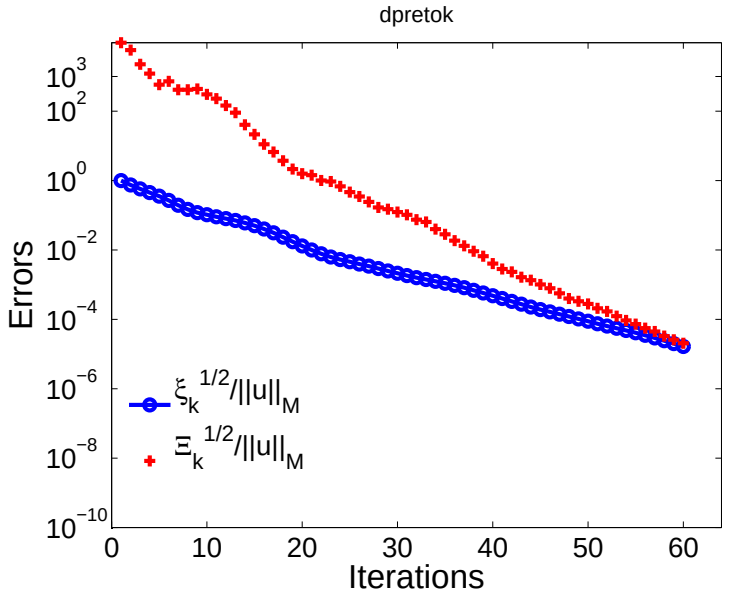
DIAMO problems Data

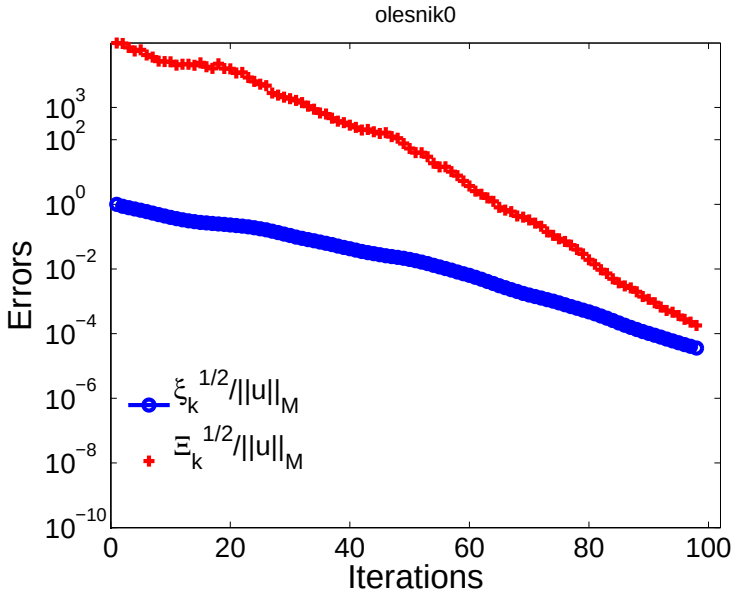
name	m	n	nnz(M)	nnz(A)	Is A full rank.
DAN2	63750	46661	220643	127054	yes
d_ pretok	129160	53570	627272	258320	no
olesnik0	61030	27233	280575	122060	no
turon	133814	56110	184158	126799	no

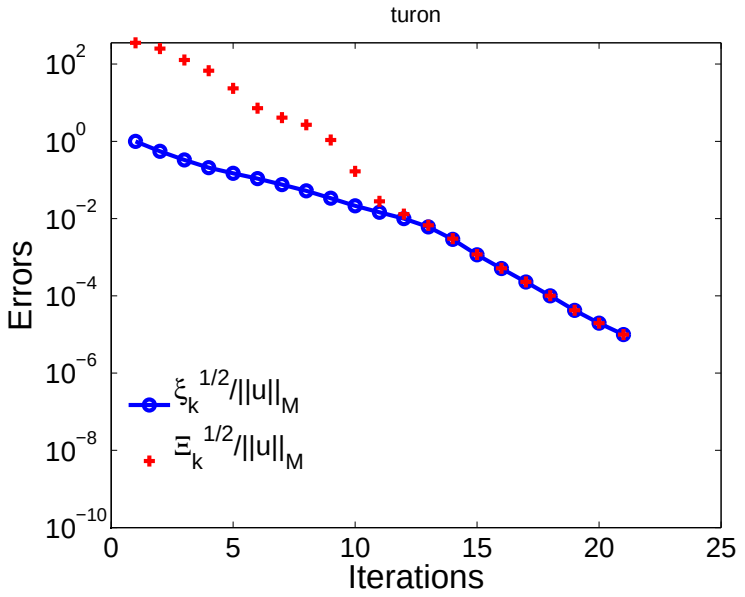
DIAMO problems data (nnz(M) is only for the symmetric part)

DIAMO problems ($d = 5$, $\tau = \frac{1}{n^2}$)









Summary and open problems

- Importance of the choice of the good norm
- Lower and upper bounds cheep and accurate (see Golub -Meurant book)
- We can prove global upper bounds of the error for mixed finite-element approximation
- How accurate must be the solution of $\mathbf{M}^{-1}\mathbf{v}$?
- How G-K can be used for solving regularized problem?

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

- How to extend the method to Banach Spaces (p-Laplacian problems)?

A. and Orban