Science & Technology
Facilities Council

Technical Report

Generalized Golub-Kahan bidiagonalization and stopping criteria

Mario Arioli

August 2, 2010

RAL-TR-2010-008

Library and Information Services
SFTC Rutherford Appleton Laboratory
Harwell Science and Innovation Campus
Didcot
OX11 0QX
UK
Tel: +44 (0)1235 445384
Fax: +44(0)1235 446403
Email: library@rl.ac.uk

The STFC ePublication archive (epubs), recording the scientific output of the Chilbolton, Daresbury, and Rutherford Appleton Laboratories is available online at:
http://epubs.cclrc.ac.uk/

ISSN 1358-6254

RAL-TR-2010-008

Generalized Golub-Kahan bidiagonalization and stopping criteria

M. Arioli[1]

The Golub-Kahan bidiagonalization algorithm has been widely used in solving least-squares problems and in the computation of the SVD of rectangular matrices. Here we propose an algorithm based on the Golub-Kahan process for the solution of augmented systems that minimizes the norm of the error and, in particular, we propose a novel estimator of the error similar to the one proposed by Hestenes-Stiefel for the conjugate gradient. This estimator gives a lower bound for the error, and can be used as a reliable stopping criterion for the whole process. We also propose an upper bound of the error base on Gauss-Radau quadrature. Finally, we show how we can transform and optimally precondition augmented systems rising from the mixed finite-element approximation of differential problems.

Keywords: Bidiagonalization, stopping criteria.

AMS(MOS) subject classifications: 65F10, 65F20, 65F50

Current reports available by anonymous ftp to ftp.numerical.rl.ac.uk in directory pub/reports.

[1] m.arioli@rl.ac.uk, Rutherford Appleton Laboratory.

The work was supported by EPSRC grant GR/S42170/01.

Computational Science and Engineering Department
Atlas Centre
Rutherford Appleton Laboratory
Oxon OX11 0QX

August 2, 2010

Contents

1 Introduction **1**

2 Generalized G-K bidiagonalization **2**

3 Stopping criteria and error estimates **5**

3.1 A lower bound estimate . . . 6

3.2 An upper bound estimate . . . 7

3.3 Mixed finite-element . . . 8

4 Numerical experiments **10**

4.1 Test problems . . . 10

4.2 Numerical results . . . 13

5 Conclusions **15**

1 Introduction

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

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

and their dual spaces

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

We remark that, using the previous notation, the matrix $\mathbf{A}$ is an operator between $\mathcal{N}$ into $\mathcal{M}$. In particular, for each fixed $\mathbf{q} \in \mathcal{N}$ we also have that

$$\langle \mathbf{v}, \mathbf{A}\mathbf{q}\rangle_{\mathcal{M},\mathcal{M}'} = \mathbf{v}^T\mathbf{A}\mathbf{q}, \quad \mathbf{A}\mathbf{q} \in \mathcal{L}(\mathcal{M}) \ \forall \mathbf{q} \in \mathcal{N}. \tag{1.1}$$

The adjoint operator $\mathbf{A}^\star$ of $\mathbf{A}$ can be defined [4] as

$$\langle \mathbf{A}^\star\mathbf{g}, \mathbf{f}\rangle_{\mathcal{N}',\mathcal{N}} = \mathbf{f}^T\mathbf{A}^T\mathbf{g}, \quad \mathbf{A}^T\mathbf{g} \in \mathcal{L}(\mathcal{N}) \ \forall \mathbf{g} \in \mathcal{M}, \tag{1.2}$$

and it is linked to the transpose of $\mathbf{A}$. Given $\mathbf{q} \in \mathcal{M}$ and $\mathbf{v} \in \mathcal{N}$, the critical points for the functional

$$\frac{\mathbf{v}^T\mathbf{A}\mathbf{q}}{\|\mathbf{q}\|_{\mathbf{N}}\,\|\mathbf{v}\|_{\mathbf{M}}} \tag{1.3}$$

are the "*generalized singular values and singular vectors*" of $\mathbf{A}$. Indeed the saddle-point conditions for (1.3) are

$$\begin{cases} \mathbf{A}\mathbf{q}_i &= \sigma_i\mathbf{M}\mathbf{v}_i \qquad \mathbf{v}_i^T\mathbf{M}\mathbf{v}_j = \delta_{ij} \\ \mathbf{A}^T\mathbf{v}_i &= \sigma_i\mathbf{N}\mathbf{q}_i \qquad \mathbf{q}_i^T\mathbf{N}\mathbf{q}_j = \delta_{ij} \end{cases} \tag{1.4}$$

Hereafter, we assume that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$. If we operate a change of variables using $\mathbf{M}^{1/2}$ and $\mathbf{N}^{1/2}$ we have that 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 1.1. *We point out that the necessary and sufficient conditions, based on the inf-sup condition [5, 7, 6], 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.*

In the following, we analyse the use of the generalized Golub-Kahan bidiagonalization algorithm (G-K bidiagonalization) [13] in solving the problem

$$\min_{\mathbf{A}^T\mathbf{u}=\mathbf{b}} \|\mathbf{u}\|^2_{\mathbf{M}}, \tag{1.5}$$

where $\mathbf{M}$ is a nonsingular symmetric and positive definite matrix.

Several problems can be reduced to the case (1.5). 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 as (1.5) by choosing

$$\left.\begin{aligned} &\mathbf{M} = \mathbf{W} + \nu\mathbf{A}\mathbf{N}^{-1}\mathbf{A}^T \\ &\mathbf{u} = \mathbf{w} - \mathbf{M}^{-1}\mathbf{g} \\ &\mathbf{b} = \mathbf{r} - \mathbf{A}^T\mathbf{M}^{-1}\mathbf{g}. \end{aligned}\right\} \tag{1.6}$$

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

The paper is organized as follows: Section 2 is dedicated to the properties of the Golub-Kahan bidiagonalization algorithm, in Section 3, we analyse and describe the stopping criteria, and finally in Section 4 we validate the theory on selected numerical examples.

2 Generalized G-K bidiagonalization

In [13, 17], 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 [3]. Here, we want specifically to analyse one of the variants known as the "Craig"-variant (see [17, 20, 19]). Therefore, we seek the upper bidiagonal matrix $\mathbf{B}$ such that the following relations are satisfied

$$\left\{\begin{array}{lclc} \mathbf{A}\mathbf{Q} & = & \mathbf{M}\mathbf{V}\begin{bmatrix}\mathbf{B}\\ 0\end{bmatrix} & \qquad \mathbf{V}^T\mathbf{M}\mathbf{V} = \mathbf{I}_m \\ \mathbf{A}^T\mathbf{V} & = & \mathbf{N}\mathbf{Q}\left[\mathbf{B}^T; 0\right] & \qquad \mathbf{Q}^T\mathbf{N}\mathbf{Q} = \mathbf{I}_n \end{array}\right. \tag{2.7}$$

where

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

The augmented system that gives the optimality conditions for (1.5)

$$\begin{bmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & 0 \end{bmatrix}\begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix} \tag{2.8}$$

can be transformed by the change of variables

$$\left\{\begin{array}{l} \mathbf{u} = \mathbf{V}\mathbf{z} \\ \mathbf{p} = \mathbf{Q}\mathbf{y} \end{array}\right. \tag{2.9}$$

and the relations (2.7) into

$$\begin{bmatrix} \mathbf{I}_n & 0 & \mathbf{B} \\ 0 & \mathbf{I}_{m-n} & 0 \\ \mathbf{B}^T & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{Q}^T\mathbf{b} \end{bmatrix}. \tag{2.10}$$

From (2.10), it possible to deduce that $\mathbf{u}$ depends only on the first n columns of $\mathbf{V}$ because $\mathbf{z}_2$ is equal to zero. Thus, we can further reduce (2.10) to the form

$$\begin{bmatrix} \mathbf{I}_n & \mathbf{B} \\ \mathbf{B}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{Q}^T\mathbf{b} \end{bmatrix}. \tag{2.11}$$

Moreover, the G-K bidiagonalization can be set such that

$$\mathbf{Q}^T\mathbf{b} = \mathbf{e}_1 \|\mathbf{b}\|_{\mathbf{N}^{-1}}$$

and, then, the value of $\mathbf{z}_1$ will correspond to the first column of the inverse of $\mathbf{B}$ multiplied by $\|\mathbf{b}\|_{\mathbf{N}^{-1}}$. By using the first relation of (2.7), we can compute the first column of $\mathbf{B}$ and of $\mathbf{V}$;

$$\alpha_1 \mathbf{M}\mathbf{v}_1 = \mathbf{A}\mathbf{q}_1, \tag{2.12}$$

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

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

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

$$\zeta_1 = \frac{\|\mathbf{b}\|_{\mathbf{N}^{-1}}}{\alpha_1}$$

as

$$\zeta_{i+1} = -\frac{\beta_i}{\alpha_{i+1}}\zeta_i \qquad i = 1, \ldots, n \tag{2.13}$$

From the first m equations of (2.8) and 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 [17, 19], we can easily transform the previous relation into a recursive one where only one extra vector is required.

First, we observe that

$$\mathbf{p}^{(k)} = -\mathbf{Q}_k\mathbf{B}_k^{-1}\mathbf{z}_k = -\left(\mathbf{B}_k^{-T}\mathbf{Q}_k^T\right)^T \mathbf{z}_k. \tag{2.14}$$

From this relation, we have that the matrix

$$\mathbf{D}_k = \mathbf{B}_k^{-T}\mathbf{Q}_k^T \tag{2.15}$$

can be computed recursively taking into account that $\mathbf{B}^T$ is a lower bidiagonal matrix as follows

$$\begin{aligned} \mathbf{d}_1 &= \frac{\mathbf{q}_1}{\alpha_1} \\ \mathbf{d}_{i+1} &= \frac{\mathbf{q}_{i+1} - \beta_{i+1}\mathbf{d}_i}{\alpha_{i+1}} \qquad i = 1, \ldots, n, \end{aligned}$$

where $\mathbf{d}_j$ are the columns of $\mathbf{D}$.

Therefore, we have that starting with $\mathbf{p}^{(1)} = -\zeta_1\mathbf{d}_1$ and $\mathbf{u}^{(1)} = \zeta_1\mathbf{v}_1$ the solutions $\mathbf{u}^{(k)}$ and $\mathbf{p}^{(k)}$ can be recursively computed as

$$\left.\begin{aligned} \mathbf{u}^{(i+1)} &= \mathbf{u}^{(i)} + \zeta_{i+1}\mathbf{v}_{i+1} \\ \mathbf{p}^{(i+1)} &= \mathbf{p}^{(i)} - \zeta_{i+1}\mathbf{d}_{i+1} \end{aligned}\right\} \qquad i = 1, \ldots, n \tag{2.16}$$

We want to point out here that the Craig algorithm we have described has an important property of minimization. Let $\mathcal{V} = span\,\{\mathbf{V}_k\}$ and $\mathcal{Q} = span\,\{\mathbf{Q}_k\}$. At each step k the algorithm 2.1 computes $\mathbf{u}^{(k)}$ such that [19]

$$\begin{array}{cc} \min & \|\mathbf{u} - \mathbf{u}^{(k)}\|_{\mathbf{M}}. \\ \mathbf{u}^{(k)} \in \mathcal{V} & \\ \left(\mathbf{A}^T\mathbf{u}^{(k)} - \mathbf{b}\right) \perp \mathcal{Q} & \end{array} \tag{2.17}$$

It is straightforward to see that the Lagrange conditions of optimality for (2.17) are satisfied from relations (2.7).

The Craig variant of G-K bidiagonalization can then be formulated as shown in Algorithm 2.1.

In the next Section 3, we give error estimates for the errors on $\mathbf{u} - \mathbf{u}^{(k)}$ and $\mathbf{p} - \mathbf{p}^{(k)}$, and on the dual norm of the residual $\mathbf{r}^{(k)} = \mathbf{A}^T\mathbf{u}^{(k)} - \mathbf{b}$.

Algorithm 2.1.

```
procedure $[\mathbf{U},\mathbf{V},\mathbf{B},\mathbf{u},\mathbf{p}]$ = G-K_bidiagonalization($\mathbf{A},\mathbf{M},\mathbf{N},\mathbf{b},maxit$);
    $\beta_1 = \|\mathbf{b}\|_{\mathbf{N}^{-1}}$; $\mathbf{q}_1 = \mathbf{N}^{-1}\mathbf{b}/\beta_1$;
    $\mathbf{w} = \mathbf{M}^{-1}\mathbf{A}\mathbf{q}_1$; $\alpha_1 = \mathbf{w}^T\mathbf{M}\mathbf{w}$; $\mathbf{v}_1 = \mathbf{w}/\sqrt{\alpha_1}$;
    $\zeta_1 = \beta_1/\alpha_1$; $\mathbf{d}_1 = \mathbf{q}_1/\alpha_1$; $\mathbf{p}^{(1)} = -\zeta_1\mathbf{d}_1$
    $k$ = 0; $it$ = 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($\mathbf{z}_k, \dots$)
    end while;
end procedure.
```

3 Stopping criteria and error estimates

Taking into account the expression of $\mathbf{u}^{(k)}$ and (2.9), we have from the $\mathbf{M}$-orthogonality properties of $\mathbf{V}$ that the $\mathbf{M}$ norm of the error $\mathbf{e}^{(k)} = \mathbf{u} - \mathbf{u}^{(k)}$ is

$$\|\mathbf{e}^{(k)}\|^2_{\mathbf{M}} = \sum_{j=k+1}^{n} \zeta_j^2 = \left\|\mathbf{z} - \begin{bmatrix} \mathbf{z}_k \\ 0 \end{bmatrix}\right\|_2^2. \tag{3.18}$$

Moreover, the dual norm of the residual

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

can easily computed. From (2.9) we have

$$\mathbf{A}^T\mathbf{u}^{(k)} - \mathbf{b} = \mathbf{A}^T\mathbf{V}\begin{bmatrix} \mathbf{z}_k \\ 0 \end{bmatrix} - \mathbf{b} = \mathbf{N}\mathbf{Q}\mathbf{B}^T\begin{bmatrix} \mathbf{z}_k \\ 0 \end{bmatrix} - \mathbf{N}\mathbf{Q}\mathbf{e}_1\|\mathbf{b}\|_{\mathbf{N}^{-1}} = \beta_{k+1}\zeta_k\mathbf{N}\mathbf{Q}\mathbf{e}_k, \tag{3.19}$$

and, thus, from (1.3) and (1.4), the dual norm is

$$\|\mathbf{A}^T\mathbf{u}^{(k)} - \mathbf{b}\|_{\mathbf{N}^{-1}} = |\beta_{k+1}\,\zeta_k| \leq \sigma_1|\zeta_k| = \|\tilde{\mathbf{A}}\|_2|\zeta_k|. \tag{3.20}$$

Finally, a bound on the $\mathbf{N}$-norm of the error $\mathbf{p} - \mathbf{p}^{(k)}$ can be obtained from (2.9) and (2.14)

$$\|\mathbf{p} - \mathbf{p}^{(k)}\|_{\mathbf{N}} = \left\|\mathbf{Q}\mathbf{B}^{-1}\left(\mathbf{z} - \begin{bmatrix} \mathbf{z}_k \\ 0 \end{bmatrix}\right)\right\|_{\mathbf{N}} \leq \frac{\|\mathbf{e}^{(k)}\|_{\mathbf{M}}}{\sigma_n}. \tag{3.21}$$

Remark 3.1. *We observe that, owing to the non singularity of both $\mathbf{M}$ and $\mathbf{N}$, all β_i and α_i, ($i = 1, \dots, n$) are strictly positive. The expression of the $\mathbf{M}$ norm of the error on $\mathbf{u}$ (3.18) and the minimization property (2.17) entail that the sequence $\|\mathbf{e}^{(k)}\|_{\mathbf{M}}$ decreases strictly.*

3.1 A lower bound estimate

From Remark 3.1, we can apply the same strategy used in [1] and proposed by Hestenes-Stiefel in [15] for testing the convergence of the conjugate gradient method. Given a threshold $\tau < 1$ and an integer d, we can estimate $\|\mathbf{e}^{(k)}\|^2_{\mathbf{M}}$ by

$$\xi^2_{k,d} = \sum_{j=k+1}^{k+d+1} \zeta_j^2 < \|\mathbf{e}^{(k)}\|^2_{\mathbf{M}}. \tag{3.22}$$

The procedure "*check*($\mathbf{z}_k, \dots$)" in Algorithm 2.1 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.
```

Although the vale of $\xi_{k,d}$ is a lower bound, it has two advantages:

- $\xi_{k,d}$ measures the error at the step $k-d$ but because all the following $\mathbf{u}^{(k)}$ minimise the error, we can safely use the last ones;
- $|\zeta_j| \leq \xi_{k,d}\ j = k-d+1, \dots, k$ and, thus, we have that $\mathbf{u}^{(k-1)}$ will satisfy

$$\|\mathbf{A}^T\mathbf{u}^{(k-1)} - \mathbf{b}\|_{\mathbf{N}^{-1}} \leq \|\tilde{\mathbf{A}}\|_2\ |\zeta_{k-1}| \leq \|\tilde{\mathbf{A}}\|_2\ \tau.$$

Furthermore, let $\mathbf{T} = \mathbf{B}^T\mathbf{B}$. $\mathbf{T}$ is a tridiagonal matrix of entries

$$\begin{cases} \mathbf{T}_{1,1} = \alpha_1^2 & \\ \mathbf{T}_{i,i} = \alpha_i^2 + \beta_i^2 & i = 2, \dots, n \\ \mathbf{T}_{i,i+1} = \mathbf{T}_{i+1,i} = \alpha_i\beta_i & i = 1, \dots, n \\ 0 & \text{otherwise} \end{cases} \tag{3.23}$$

and $\mathbf{T}$ is a nonnegative positive definite matrix. The $\|\mathbf{z}\|_2^2$ is then equal to

$$\|\mathbf{b}\|^2_{\mathbf{N}^{-1}}\ \left(\mathbf{T}^{-1}\right)_{1,1}$$

the $(1,1)$ entry of the inverse of $\mathbf{T}$. Thus, we have

$$\|\mathbf{e}^{(k)}\|^2_{\mathbf{M}} = \sum_{j=k+1}^{n} \zeta_j^2 = \|b\|^2_{\mathbf{N}^{-1}} \left[\left(\mathbf{T}^{-1}\right)_{1,1} - \left(\mathbf{T}_k^{-1}\right)_{1,1}\right], \tag{3.24}$$

where $\mathbf{T}_k$ is the $k \times k$ principal submatrix of $\mathbf{T}$.

Following [14] exposition, the estimate of the error can then be interpreted as a Gaussian quadrature process for approximating the integral of a function with respect to a Stieltjes measure defined by the singular values of $\mathbf{B}$. In [14, Ch. 6], it is shown as this gives an estimate that is a lower bound of the exact integral.

3.2 An upper bound estimate

Despite being very inexpensive, the estimator (3.22) is still a lower bound of the error. It would also be useful to have an upper bound estimator of the error. Taking into account the observation at the end of the previous section, we can use an approach inspired by the Gauss-Radau quadrature algorithm and similar to the one described in [14, Ch. 6].

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

$$\hat{\mathbf{T}}_{k+1} = \begin{bmatrix} \mathbf{T}_k & \alpha_k\beta_k\mathbf{e}_k \\ \alpha_k\beta_k\mathbf{e}_k^T & \omega_{k+1} \end{bmatrix}, \tag{3.25}$$

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

$$\left(\mathbf{T}_k - a^2\mathbf{I}\right)\delta(a^2) = \alpha_k^2\beta_k^2\mathbf{e}_k. \tag{3.26}$$

We point out that the matrix $\left(\mathbf{T}_k - a^2\mathbf{I}\right)$ is positive definite and that $\hat{\mathbf{T}}_{k+1}$ has one eigenvalue equal to a^2.

Analogously to what is done in [14] for the conjugate gradient method, we can recursively compute $\delta(a^2)_k$ and ω_{k+1} by using the Cholesky decomposition. Therefore, we obtain the following realization of the procedure "*check*($\mathbf{z}_k, \dots$)" in Algorithm 2.1;

```
procedure [ convergence ] = checkGR($\mathbf{z}_k, k, d, \tau, a, \|b\|_{\mathbf{N}^{-1}}, \mathbf{B}_k$)
    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;
    $\varpi_k = a^2 + \frac{\alpha_k^2\beta_k^2}{\bar{d}_k}$;   $\varphi_k = \frac{\beta_k^2\zeta_k^2}{\sqrt{\bar{d}_k + a^2 - \beta_k^2}}$;

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

        if $\bar{\xi} \le \tau$ then;
            convergence = true;
        end if;

    end if;
end procedure.
```

The procedure "checkGR" is the practical realization of a Gauss-Radau quadrature that uses the matrices $\hat{\mathbf{T}}_k$. Therefore, from [14, Theorem 6.4], we can derive that $\bar{\xi}$ is an upper bound for $\|\mathbf{e}^{(k)}\|_{\mathbf{M}}$.

The major drawback of the Gauss-Radau approach is the need for an accurate estimate of the smallest singular value of $\mathbf{B}$. This can be very difficult in general, however, in special cases, it can be done. In particular, taking into account Remark 1.1, if the problem is the approximation of a variational problem by mixed (hybrid) finite-element methods, and the *inf-sup* condition is satisfied [5, 7] a good estimate can be derived from the physical properties of the underlying continuous model. In [14], adaptive strategies are proposed in order to improve the estimate of a.

Another difficulty arises from the choice of the delay d. Again, this is very much problem dependent and $\mathbf{M}$ and $\mathbf{N}$ dependent. If $\mathbf{M}$ and $\mathbf{N}$ can be chosen such that the generalized singular values of $\mathbf{A}$ become bounded in an interval independent of n and m, the delay parameter d can be quite small.

In the next section, we analyse the connections of the proposed stopping criteria for the mixed/hybrid finite element method.

Remark 3.2. *We presented here only the upper bound for the Gauss-Radau for the sake of simplicity. As illustrated in [14], it is possible to obtain another upper bound based on Gauss-Lobatto integration theory. In this case, we need both a lower bound a and an upper bound b for the singular values of $\mathbf{B}$, i.e. $a \leq \sigma_n$ and $\sigma_1 \leq b$.*

3.3 Mixed finite-element

In this section, we apply the results of Section 3.1 and Section 3.2 to the solution of continuous saddle-point problem [5, 7]. The aim is to have error bounds merging the approximation error for the mixed finite-element method and the algebraic errors introduced by the generalized G-K bidiagonalization method. Let $\mathcal{H}$ and $\mathcal{P}$ be two Hilbert spaces, and $\mathcal{H}'$ and $\mathcal{P}'$ the corresponding dual spaces. Let

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

be continuous bilinear forms. Given $f \in \mathcal{H}'$ and $g \in \mathcal{P}'$, we seek the solutions $u \in \mathcal{H}$ and $p \in \mathcal{P}$ of the system

$$\begin{aligned}
\mathfrak{a}(u,v) + \mathfrak{b}(v,p) \quad &= \langle f, v \rangle_{\mathcal{H}',\mathcal{H}} \quad \forall v \in \mathcal{H} \\
\mathfrak{b}(u,q) \quad &= \langle g, q \rangle_{\mathcal{P}',\mathcal{P}} \quad \forall q \in \mathcal{P}.
\end{aligned} \tag{3.27}$$

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

$$\begin{aligned}
M \quad &: \quad \mathcal{H} \to \mathcal{H}', \quad \langle Mu, v \rangle_{\mathcal{H}' \times \mathcal{H}} = \mathfrak{a}(u,v) \quad \forall u \in \mathcal{H}, \forall v \in \mathcal{H} \\
A^\star \quad &: \quad \mathcal{H} \to \mathcal{P}', \quad \langle A^\star u, q \rangle_{\mathcal{P}' \times \mathcal{P}} = \mathfrak{b}(u,q) \quad \forall u \in \mathcal{H}, \forall q \in \mathcal{P} \\
A \quad &: \quad \mathcal{P} \to \mathcal{H}', \quad \langle u, Aq \rangle_{\mathcal{H} \times \mathcal{H}'} = \mathfrak{b}(u,q) \quad \forall u \in \mathcal{H}, \forall q \in \mathcal{P}
\end{aligned}$$

and we have

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

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

$$0 < \chi_1 \|v\|_{\mathcal{H}} \leq \mathfrak{a}(u,u). \tag{3.28}$$

However, [7] the coercivity on the $Ker(A^\bigstar)$ is sufficient. We will also assume that $\exists\chi_0 > 0$ such that

$$\sup_{v\in\mathcal{H}} \frac{\mathfrak{b}(v,q)}{\|v\|_{\mathcal{H}}} \geq \chi_0 \|q\|_{\mathcal{P}\setminus Ker(A)}. \tag{3.29}$$

Under the hypotheses (3.28), (3.29), and for any $f \in \mathcal{H}'$ and $g \in Im(A^\bigstar)$ then there exists (u, p) solution of (3.27) [7, Theorem 1.1]. Moreover, (see [7, Theorem 1.1]) u is unique and p is definite up to an element of $Ker(A)$.

Let now $\mathcal{H}_h \hookrightarrow \mathcal{H}$ and $\mathcal{P}_h \hookrightarrow \mathcal{P}$ be two finite dimensional subspaces of $\mathcal{H}$ and $\mathcal{P}$. As for the problem (3.27), we can introduce the operators $A_h : \mathcal{P}_h \to \mathcal{H}'_h$ and $M_h; \mathcal{H}_h \to \mathcal{H}'_h$. We also assume that

$$\begin{cases} Ker(A_h) \subset Ker(A) \\ \sup_{v_h\in\mathcal{H}_h} \dfrac{\mathfrak{b}(v_u, q_h)}{\|v_h\|_{\mathcal{H}}} \geq \chi_n \|q_h\|_{\mathcal{P}\setminus Ker(A_h)} \\ \chi_n \geq \chi_0 > 0. \end{cases} \tag{3.30}$$

Under the hypotheses (3.28), (3.29), and (3.30), ([7, Proposition 2.1 and Theorem 2.1]), we have that $\exists(u_h, p_h) \in \mathcal{H}_h \times \mathcal{P}_h$ solution of

$$\begin{aligned} \mathfrak{a}(u_h, v_h) + \mathfrak{b}(v_h, p_h) &= \langle f, v_h\rangle_{\mathcal{H}'_h,\mathcal{H}_h} \quad \forall v_h \in \mathcal{H}_h \\ \mathfrak{b}(u_h, q_h) &= \langle g, q_h\rangle_{\mathcal{P}'_h,\mathcal{P}_h} \quad \forall q_h \in \mathcal{P}_h. \end{aligned} \tag{3.31}$$

and

$$\begin{aligned} \|u - u_h\|_{\mathcal{H}} \quad + \quad & \|p - p_h\|_{\mathcal{P}\setminus Ker(A)} \leq \\ & \kappa \left(\inf_{v_h\in\mathcal{H}_h} \|u - v_h\|_{\mathcal{H}} + \inf_{q_h\in\mathcal{P}_h} \|p - q_h\|_{\mathcal{P}} \right), \end{aligned} \tag{3.32}$$

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

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

- s generalized singular values will be zero;
- however, the G-K bidiagonalization method will still work and it will compute a matrix $\mathbf{B}$ of rank $n - s$.

On the basis of the latter observations, the error $\|\mathbf{e}^{(k)}\|_{\mathbf{M}}$ can be still computed by (3.18) and the bounds (3.20) and (3.21) hold. Finally, we point out the (3.30) imply that for $h \downarrow 0$ the generalized singular values of all $\mathbf{A} \in \mathbb{R}^{m_h \times n_h}$ will be bounded with upper and lower bounds independent of h, i.e.

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

We can then prove the following Theorem.

Theorem 3.1. *Under (3.28), (3.29), and (3.30), and denoting by* $\mathbf{u}^*$ *one of the iterates of Algorithm 2.1 for which* $\|\mathbf{e}^{(k)}\|_{\mathbf{M}} < \tau$*, we have*

$$\begin{aligned} \|u(x) - u^*(x)\|_{\mathcal{H}} \quad + \quad & \|p - p^*(x)\|_{\mathcal{P}\setminus Ker(A)} \leq \\ & \breve{\kappa} \left(\inf_{v_h\in\mathcal{H}_h} \|u - v_h\|_{\mathcal{H}} + \inf_{q_h\in\mathcal{P}_h} \|p - q_h\|_{\mathcal{P}} + \tau \right), \end{aligned} \tag{3.33}$$

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

Proof.

$$\begin{aligned}\|u(x) - u^*(x)\|_{\mathcal{H}} \quad + & \quad \|p - p^*(x)\|_{\mathcal{P}\backslash Ker(A)} \leq \\ & \quad \|u(x) - u_h\|_{\mathcal{H}} + \|p - p_h\|_{\mathcal{P}\backslash Ker(A)} + \\ & \quad \|u_h - u^*(x)\|_{\mathcal{H}} + \|p_h - p^*(x)\|_{\mathcal{P}\backslash Ker(A)}.\end{aligned}$$

From (3.21), we have

$$\|u_h - u^*(x)\|_{\mathcal{H}} + \|p_h - p^*(x)\|_{\mathcal{P}\backslash Ker(A)} \leq \left(1 + \frac{1}{\sigma_n}\right) \|\mathbf{e}^{(k)}\|_{\mathbf{M}},$$

and from (3.29)

$$\left(1 + \frac{1}{\sigma_n}\right) \leq \left(1 + \frac{1}{\chi_0}\right).$$

Thus, (3.33) follows with

$$\breve{\kappa} = \max\left(\kappa, \left(1 + \frac{1}{\chi_0}\right)\right).$$

□

The inequality (3.33) gives an easy way to choose the threshold τ. In practice, separate upper bounds for

$$\inf_{v_h \in \mathcal{H}_h} \|u - v_h\|_{\mathcal{H}}, \qquad \text{and} \qquad \inf_{q_h \in \mathcal{P}_h} \|p - q_h\|_{\mathcal{P}},$$

can be obtained both a priori [7] or a posteriori [16], and τ can be chosen as a scalar of the same order.

Finally, we point out that the Algorithm 2.1 convergence rate will not depend on the dimensions of the problem.

4 Numerical experiments

4.1 Test problems

We have four classes of test problems:

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

$$\begin{aligned} -\nabla \cdot \nabla u &= f \qquad \text{in } \Omega, & (4.34)\\ \frac{\partial u}{\partial \mathbf{n}} &= 0 \qquad \text{on } \partial_N \Omega = \{0 \times (0, 1)\} \cup \{1 \times (0, 1)\}, & (4.35)\\ u &= 0 \qquad \text{on } \partial_D \Omega = \{(0, 1) \times 0\} & (4.36)\\ u &= 1 \qquad \text{on } \partial_D \Omega = \{(0, 1) \times 1\}. & (4.37)\end{aligned}$$

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

- The Poisson equation with Neumann zero boundary conditions [18] on a domain $\Omega = (0, 1) \times (0, 1)$:

$$\begin{aligned} -\nabla \cdot \nabla u &= f \qquad \text{in } \Omega, & (4.38)\\ \frac{\partial u}{\partial \mathbf{n}} &= 0 \qquad \text{on } \partial \Omega & (4.39)\end{aligned}$$

  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, that represent fluid flow in porous media: these test problems [8] can be downloaded from http://www.cise.ufl.edu/research/sparse/matrices.

Poisson with mixed finite-element approximation Following [7], the Poisson problem is casted in its dual form as a Darcy's problem:

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

We subdivided Ω with a uniform mesh of triangles, see Figure 4.1. Then, we approximated the spaces $\mathcal{H}$ and $L^2(\Omega)$ by RT0 and by piecewise constant functions respectively using the Matlab software described in [2]. 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

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

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

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

We point out that the pattern of $\mathbf{W}$ is structurally equal to the pattern $\mathbf{A}\mathbf{N}^{-1}\mathbf{A}^T$. In Table 4.1, we give the dimensions, the value of h, the number of nonzero entries in $\mathbf{A}$ and in the upper triangular part of $\mathbf{M}$ for each generated mesh.

Moreover, the discrete norm of $\mathbf{q}$ is equal to the $\mathcal{H}$ norm of the corresponding finite dimensional function given by the linear combination of the basis functions with $\mathbf{q}$. With the chosen boundary conditions, it is easy to verify that the continuous solution u is $u(x, y) = x$.

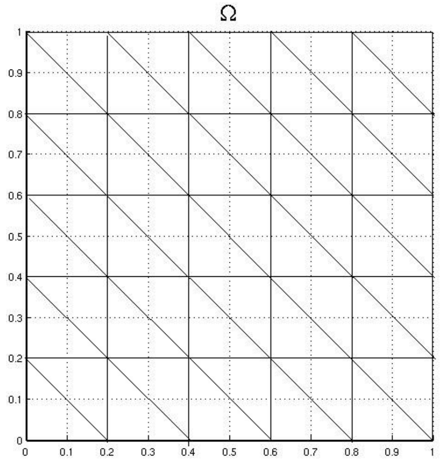


Figure 4.1: An example of uniform triangulation

$h = 2^{-k}$	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

Table 4.1: Poisson with mixed b.c.data and RT0 (nnz($\mathbf{M}$) is only for the symmetric part)

Neumann problem Following [18], we introduce the function $\vec{w}(x) = -\nabla u$ and, then we rewrite equation (4.38) in Darcy form

$$\left.\begin{array}{rcl} \vec{w}(x) + \nabla u & = & 0 \\ \nabla \cdot \vec{w}(x) & = & f \end{array}\right\}. \tag{4.40}$$

As in [18], 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. After scaling, the finite dimensional problem has the following structure:

$$\begin{bmatrix} \mathbf{I}_m & \mathbf{E} \\ \mathbf{E}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix}.$$

Given the bidiagonal matrix $\mathbf{C} \in \mathbb{R}^{\sqrt{n}-1 \times \sqrt{n}}$

$$\mathbf{C}_{i,j} = \begin{cases} -1 & i = j, \\ 1 & j = i+1, \\ 0 & \text{otherwise}, \end{cases}$$

the matrix $\mathbf{E}$ is

$$\mathbf{E} = \frac{1}{\sqrt{n}-1} \begin{bmatrix} \mathbf{I}_{\sqrt{n}} \otimes \mathbf{C} \\ \mathbf{C} \otimes \mathbf{I}_{\sqrt{n}} \end{bmatrix},$$

and the vector $\mathbf{b}$ has entries equal to the values of the function f in the nodes divided by $\sqrt{n}$.

We point out that the matrix $\mathbf{E}$ is not full rank (the product of $\mathbf{E}$ by a vector with all equal entries produces the zero vector). 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

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

In all test problems, we modified the $(1,1)$ block adding to it the matrix $\mathbf{E}\mathbf{E}^T$ and, then, corrected the right-hand side consistently with (1.6). In Table 4.2, we give the dimensions, the number of nonzero entries in $\mathbf{A}$ and in the upper triangular part of $\mathbf{M}$ for each generated mesh.

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

Table 4.2: Poisson with Neumann b.c. data (nnz($\mathbf{M}$) is only for the symmetric part)

Stokes problems The Stokes problems have been generated using the software provided by **ifiss3.0** package [9, 10]. We use the default geometry of "Step case" and the **Q2-Q1** approximation described in [22, page 27]. In Table 4.3, we give the dimensions, the number of nonzero entries in $\mathbf{A}$ and in the upper triangular part of $\mathbf{M}$ for each generated mesh.

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

Table 4.3: Stokes problems data (nnz($\mathbf{M}$) is only for the symmetric part)

DIAMO problems The DIAMO problems have been downloaded. As for the Poisson problem, we added to the (1,1) block the matrix $\mathbf{A}\mathbf{A}^T$ and modified the right-hand side following (1.6). We have the original right-hand side only for the DEN2 problem. For all the remaining ones, we use a value of $\mathbf{b}$ similar to the one used for the Poisson problems.

In Table 4.4, we give the dimensions, the number of nonzero entries in $\mathbf{A}$ and in the upper triangular part of $\mathbf{M}$ for each generated mesh. Moreover, we indicate if $\mathbf{A}$ has full rank.

4.2 Numerical results

For all the experiments, we report the summary of the results obtained using a Matlab version of Algorithm 2.1, where the matrices $\mathbf{M}$ and $\mathbf{N}$ are factorized using the Matlab function **chol** ($[R, err, S] = chol(\mathbf{X})$, $\mathbf{X} = \mathbf{M}$ or $\mathbf{Q}$). For the dual formulation of the Poisson problem (4.34), we chose $f = 0$ and, thus we could compare the error on the computed solution exactly. In Table 4.5, we display the number of iterations, $\|\mathbf{e}^{(k)}\|_2$ for the velocity field, the residual $\mathbf{A}^T\mathbf{u} - \mathbf{b}$, and $\|\mathbf{p} - \mathbf{p}^{(k)}\|_2$ for $\tau = 10^{-8}$. The number of steps is independent of the value of h. The number of iterations in the table includes the extra d steps.

name	m	n	nnz($\mathbf{M}$)	nnz($\mathbf{A}$)	Is $\mathbf{A}$ 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

Table 4.4: DIAMO problems data (nnz($\mathbf{M}$) is only for the symmetric part)

$h = 2^{-k}$	# 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

Table 4.5: Poisson with mixed b.c. data and RT0 problem results ($d = 5$, $\tau = 10^{-8}$).

In Figure 4.2, we display the convergence behaviour for the problem corresponding to $h = 2^{-8}$.

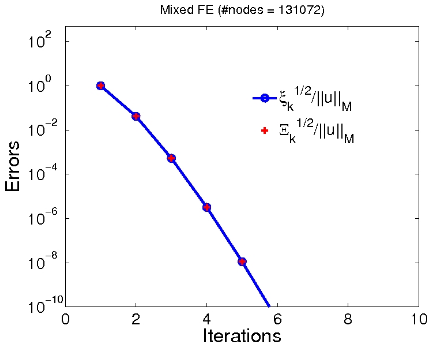


Figure 4.2: Convergence behaviour for Poisson with mixed b.c. data and RT0 problem ($h = 2^{-8}$, $d = 5$, $\tau = 10^{-8}$).

For the Stokes problems, we have a well defined Hilbert space $\mathcal{H}$ for the velocity filed:

$$\mathcal{H} = H^1(\Omega) \times H^1(\Omega).$$

Moreover, the pressure (or potential) is defined in $\mathcal{P} = L^2(\Omega)$. The package **ifiss3.0** gives the possibility of supplying the mass matrix $\mathbf{N}$ for the approximation of the norm on $L^2(\Omega)$. Thus, we can compute the generalized G-K bidiagonalization using the correct approximation of the norms for $\mathcal{H}$ and $\mathcal{P}$. It is not surprising that, in agreement with the results of [24, 21, 10], Algorithm 2.1 computes the solutions of the problem in a number of steps independent of the size of the mesh.

In Table 4.2, we report the summary of the results, where the tolerance τ was fixed at $\tau = 10^{-10}$ and the value of the delay d was five. The number of iterations in the table includes the extra d steps. On this set of problems the Gauss-Radau upper bounds are very close to the lower bounds. In Figure 4.2, we display the convergence behaviour for problem Step5.

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

Table 4.6: Stokes problems results ($d = 5$, $\tau = 10^{-10}$).

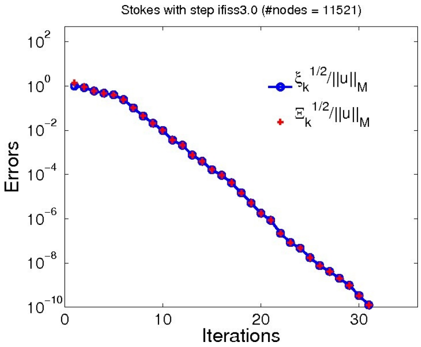


Figure 4.3: Convergence behaviour for problem Step5 ($d = 5$, $\tau = 10^{-10}$).

In Table 4.2, we display the results for the Neumann problems with $\tau = 10^{-8}$ and $d = 5$. For these problems, the value of the pressure is known minus a constant and we do not give the true error value in the table. The convergence rate of Algorithm 2.1 is independent of n in these problems. The number of iterations in the table includes the extra d steps. In Figure 4.4, we display the convergence behaviour for problem NFD5.

Finally, in Figure 4.5, we display the convergence behaviours for the DIAMO set of problems. In these cases it was impossible to have a reasonable evaluation of the norm of $\mathcal{P}$ and we used $1/n^2$ by the identity of order m as an approximation of the norm. In all the cases, the value of τ has been chosen as $\tau = \frac{1}{n^2}$.

5 Conclusions

Several authors have discussed the relations between G-K algorithm and Minres algorithm [17, 3, 19]. Moreover, the use of block diagonal preconditioners as optimal preconditioners for Krylov methods has been proved in [21, 10] for systems arising in the solution of fluid-dynamic problems. It is known that Minres can exhibit some form of stagnation every other step, i.e. after one step where the global residual of the system decreases, the successive step does not.

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

Table 4.7: Poisson with Neumann b.c. problems: results ($d = 5$, $\tau = 10^{-8}$).

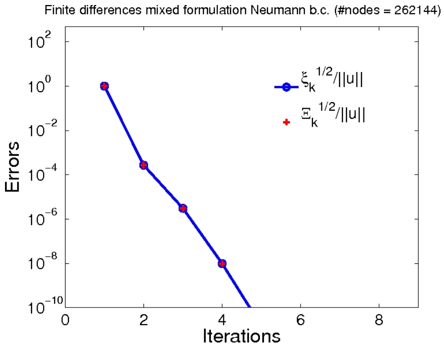


Figure 4.4: Convergence behaviour for problem NFD5 ($d = 5$, $\tau = 10^{-8}$).

Here, we have generalized the Craig version of G-K bidiagonalization to the augmented system. Taking advantage of the inf-sup conditions, we have shown that the problem can be transformed into an equivalent one where the proposed algorithm converges with a rate independent of the dimension of the problem itself.

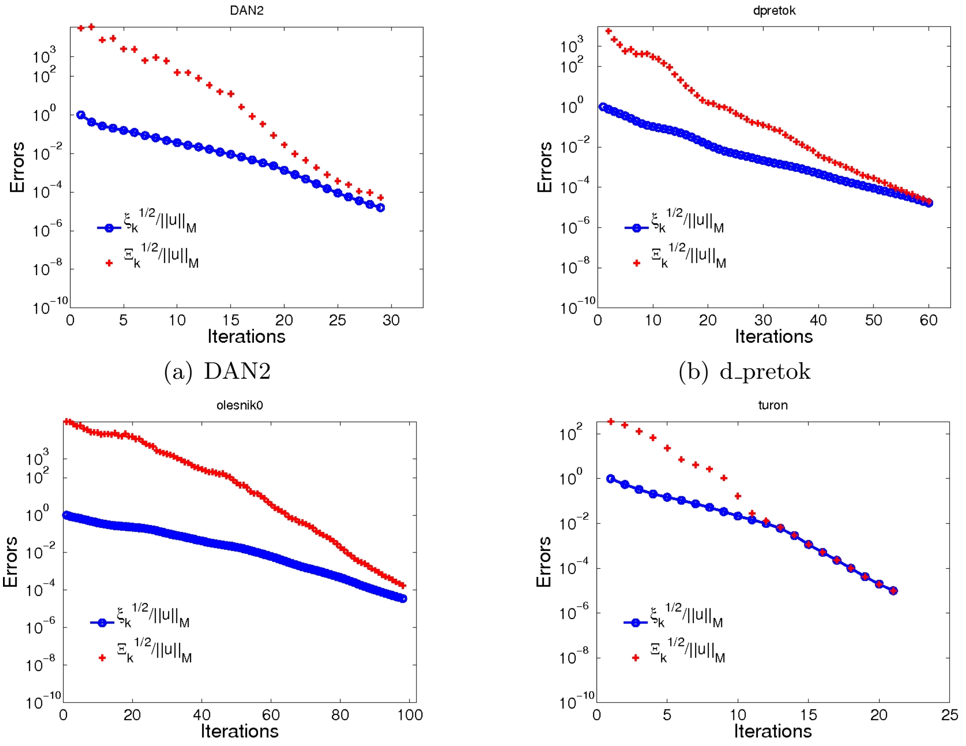
We point out that the cost of one iteration of Algorithm 2.1 is the same of one step of Minres with a block diagonal preconditioner made with $\mathbf{M}$ and $\mathbf{N}$. At each step the error $\|\mathbf{e}^{(k)}\|^2_{\mathbf{M}}$ decreases and the additional steps necessary to estimate the convergence can be very moderate if we choose the correct norms.

We proposed both a lower and an upper bound estimator of the error that are reliable and computationally inexpensive. Their properties follow easily from the results of [14]. The upper bound can be computationally problematic for problems where the matrix $\mathbf{A}$ is rank deficient as we have noticed in solving the Neumann test problems. However, the lower bound has been always very accurate.

Finally, we observe that Algorithm 2.1 cannot be extended straightforwardly to the stabilized version of augmented systems such as

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

The matrix governing the system is now *symmetric quasi-definite* [11] and in this case an LDL^T factorization without pivot can be a viable alternative [12, 23]. Moreover, in this case, it would be appropriate to investigate the use of a conjugate gradient-like algorithm.



(a) DAN2

(b) d_pretok

(c) olensk0

(d) turon

Figure 4.5: DIAMO problems results ($d = 5$, $\tau = \dfrac{1}{n^2}$).

References

[1] M. Arioli, *A stopping criterion for the conjugate gradient algorithm in a finite element method framework*, Numer. Math., 97 (2004), pp. 1–24. Electronic version: DOI: 10.1007/s00211-003-0500-y.

[2] C. Bahriawati and C. Carstensen, *Three Matlab implementations of the lowest-order Raviart-Thomas MFEM with a posteriori error control*, Computational Methods In Applied Mathematics, 5 (2005), pp. 333–361.

[3] S. J. Benbow, *Solving generalized least-squares problems with LSQR*, SIAM Journal on Matrix Analysis and Applications, 21 (1999), pp. 166–177.

[4] H. Brézis, *Analyse fonctionnelle : théorie et applications*, Dunod, Paris, 1983.

[5] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from lagrangian multipliers*, R.A.I.R.O., 8 (1974), pp. 129–151.

[6] ——, *Stability of saddle-points in finite dimensions*, in Frontiers in Numerical Analysis, J. F. Blowey, A. W. Craig, and T. Shardlow, eds., Universitext, 2003, pp. 17–62.

[7] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New-York, 1991.

[8] T. A. Davis, *The University of Florida Sparse Matrix Collection*, tech. rep., University of Florida, 2010. Submitted to ACM Transactions on Mathematical Software.

[9] H. C. Elman, A. Ramage, and D. J. Silvester, *Algorithm 866: Ifiss, a Matlab toolbox for modelling incompressible flow*, ACM Trans. Math. Softw., 33 (2007), p. 14.

[10] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers with Applications in Incompressible Fluid Dynamics*, Numerical Mathematics and Scientific Computation, Oxfgord University Press, 2005.

[11] A. George and K. Ikramov, *On the condition of symmetric quasi-definite matrices*, SIAM Journal on Matrix Analysis and Applications, 21 (2000), pp. 970–977.

[12] P. E. Gill, M. A. Saunders, and J. R. Shinnerl, *On the stability of Cholesky factorization for symmetric quasidefinite systems*, SIAM Journal on Optimization, 17 (1996), pp. 35–46.

[13] G. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis, 2 (1965), pp. 205–224.

[14] G. H. Golub and G. Meurant, *Matrices, Moments and Quadrature with Applications*, Princeton University Press, 2010.

[15] M. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–436.

[16] P. Jiránek, Z. Strakoš, and M. Vohralík, *A posteriori error estimates including algebraic error and stopping criteria for iterative solvers*, to appear on SISC, (2010).

[17] C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Transactions on Mathematical Software, 8 (1982), pp. 43–71.

[18] V. Sarin and A. Sameh, *An efficient iterative method for the generalized Stokes problem*, SIAM Journal on Scientific Computing, 19 (1998), pp. 206–226.

[19] M. Saunders, *Solution of sparse rectangular systems using LSQR and Craig*, BIT, 35 (1995).

[20] ——, *Computing projections with LSQR*, BIT, 37 (1997), pp. 96–104.

[21] D. Silvester and A. Wathen, *Fast iterative solution of stabilised Stokes systems. Part II: Using general block preconditioners*, SIAM Journal on Numerical Analysis, 31 (1994), pp. 1352–1367.

[22] D. J. Silvester, H. C. Elman, and A. Ramage, *Incompressible Flow and Iterative Solver Software Version 3.0*, Manchester University, http://www.maths.manchester.ac.uk/ djs/ifiss/, version 3.0 ed., 2009.

[23] R. J. Vanderbei, *Symmetric quasi-definite matrices*, SIAM Journal on Optimization, 5 (1995), pp. 100–113.

[24] A. Wathen and D. Silvester, *Fast iterative solution of stabilised Stokes systems. Part I: Using simple diagonal preconditioners*, SIAM Journal on Numerical Analysis, 30 (1993), pp. 630–649.