Technical Report
RAL-TR-2006-007

CCLRC

A note on GMRES preconditioned by a perturbed LDL^T decomposition with static pivoting

Mario Arioli, Iain S. Duff, Serge Gratton, and Stéphane Pralet

March 13, 2007

Council for the Central Laboratory of the Research Councils

Library and Information Services
CCLRC Rutherford Appleton Laboratory
Chilton Didcot
Oxfordshire OX11 0QX
UK
Tel: +44 (0)1235 445384
Fax: +44(0)1235 446403
Email: library@rl.ac.uk

CCLRC reports are available online at:
http://www.clrc.ac.uk/Activity/ACTIVITY=Publications;SECTION=225;

ISSN 1358-6254

RAL-TR-2006-007

(Revised)

A note on GMRES preconditioned by a perturbed LDL^T decomposition with static pivoting

M. Arioli[1], I. S. Duff[1,2], S. Gratton[2], and S. Pralet[3]

ABSTRACT

A strict adherence to threshold pivoting in the direct solution of symmetric indefinite problems can result in substantially more work and storage than forecast by an sparse analysis of the symmetric problem. One way of avoiding this is to use static pivoting where the data structures and pivoting sequence generated by the analysis are respected and pivots that would otherwise be very small are replaced by a user defined quantity. This can give a stable factorization but of a perturbed matrix.

The conventional way of solving the sparse linear system is then to use iterative refinement (IR) but there are cases where this fails to converge. In this paper, we discuss the use of more robust iterative methods, namely GMRES and FGMRES.

We show both theoretically and experimentally that both these approaches are more robust than IR and furthermore that FGMRES is far more robust than GMRES and that, under reasonable hypotheses, FGMRES is backward stable. We also show how restarted variants can be beneficial although again the GMRES variant is not as robust as FGMRES.

Keywords: augmented systems, sparse matrices, GMRES, flexible GMRES, static pivoting, roundoff error.

AMS(MOS) subject classifications: 65F05, 65F50, 65F10, 65G50

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

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

[2] Serge.Gratton@cerfacs.fr, 42, CERFACS, rue G. Coriolis, 31057 Toulouse, France

[3] stephane.pralet@samcef.com.SAMTECH s.a. LIEGE Science Park, Rue des Chasseurs-Ardennais, 8, B-4031 Liège, Belgium. Part of the work of this author was supported by ENSEEIHT and CNRS.
The work of first two authors was supported by EPSRC grant GR/S42170/01.

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

March 13, 2007

Contents

1 Introduction

This paper is concerned with solving the set of linear equations

$$Ax = b. \tag{1.1}$$

where the coefficient matrix $A \in \mathbb{R}^{n\times n}$ is a symmetric indefinite sparse matrix. Our hope is to solve this system using a direct method but sometimes the cost of doing this is too high in terms of time or memory. We have therefore looked at the possibility of using static pivoting to avoid these problems which are particularly acute if the matrix is highly indefinite as for example can happen for saddle-point problems.

We will use a multifrontal approach for our direct method. In this approach, we first determine an order for choosing pivots based on the sparsity structure of A, and we then accommodate further pivoting for numerical stability during the subsequent numerical factorization phase. When the matrix is highly indefinite, the resulting pivot sequence used in the numerical factorization can differ substantially from that predicted by the analysis step. A simple way to avoid this problem is to force the elimination through static pivoting.

We assume that the matrix A has been factorized using the HSL `MA57` package (Duff 2004) with the option of using static pivoting (Duff and Pralet 2005). Static pivoting will add a factor τ to a diagonal entry when it is impossible to find a suitable pivot in the fully summed blocks. It is common to choose $\tau \approx \sqrt{\varepsilon}||A||$ (ε machine precision).

Therefore, the computed factors $\hat{L}$ and $\hat{D}$ are, in exact arithmetic, the exact factorization of the perturbed problem

$$A + E = \hat{L}\hat{D}\hat{L}^T, \tag{1.2}$$

where E is a diagonal matrix of rank equal to the number of static pivots used during the factorization and $|E| \leq \tau I$. Hereafter, given the matrix B, we denote by $|B|$ the matrix having entries equal to the absolute values of the entries of B. The nonzero diagonal entries in E correspond to the positions at which static pivoting was performed and they are all equal to τ. Note that if τ is chosen too small then the factorization could be very unstable whereas if it is chosen too large, the factorization will be stable but will not be an accurate factorization of the original matrix (that is, $|E|$ will be large). In Section 2, we give a brief description of the static pivoting strategy used in `MA57` .

Equation (1.2) gives a splitting of A in terms of $M = \hat{L}\hat{D}\hat{L}^T$ and E

$$A = M - E, \tag{1.3}$$

and the solution of (1.1) can be expressed as the solution of the equivalent systems

$$\begin{cases} (I - EM^{-1})Mx = b \\ (I - M^{-1}E)x = M^{-1}b. \end{cases} \tag{1.4}$$

Owing to the symmetry of all the matrices in (1.4), $I - EM^{-1} = (I - M^{-1}E)^T$. Moreover, the first system corresponds to a right preconditioning of (1.1), while the second corresponds to a left preconditioning.

If the spectral radius of the matrix $I - M^{-1}E$ (or $I - EM^{-1}$) is less than one, the system (1.4) can be solved using iterative refinement. This has been used by many authors (Demmel, Hida, Kahan, Li, Mukherjee and Riedy 2005, Duff and Pralet 2005, Higham 2002, Skeel 1980) and is successful over a wide range of matrices although it is sensitive to the value of τ. If, however,

the spectral radius is greater than or equal to one (or even close to one), it may be necessary to switch to a more powerful method like GMRES (Saad and Schultz 1986). Although the matrix is symmetric, we choose GMRES since it gives us much more freedom to work with a wide range of preprocessors and preconditionings.

We have found experimentally that using the factorization (1.2) as a right preconditioning for GMRES works in most cases and is, as expected, much more robust than iterative refinement. However, there are cases where we do not get convergence of a scaled residual to machine precision. Moreover, we have found that restarted GMRES performs better and that using FGMRES (Saad 1993), even though our preconditioner remains constant, does even better.

We illustrate this through numerical experiment (see Section 6) and show theoretically that, under reasonable assumptions, FGMRES preconditioned by our static pivoting factorization is backward stable so that a small scaled residual can be achieved (see Section 5). Our analysis also holds for the case of restarted FGMRES that we advocate for controlling the memory requirement while still achieving the desired robustness and accuracy. Indeed we give theoretical arguments why restarting often greatly improves the convergence.

In the following, we use the MATLAB notation $[A, B]$ to mean the matrix that consists of the columns of matrix A followed by the columns of matrix B. We denote by ε the machine precision of a finite-precision arithmetic satisfying the IEEE standard, and by $fl(\cdot)$ the result of a sequence of floating-point operations. The computed matrices, vectors, and scalars will be identified by a bar over the symbol. Finally, we will denote by $||\cdot||$ the usual Euclidean 2-norm for vector and the corresponding induced norm for matrices.

2 $L^T DL$ with static pivoting

In the multifrontal context, the factorization can be represented by a tree at each node of which elimination operations are performed on a partially summed frontal matrix

$$\begin{bmatrix} F_{11} & F_{12} \\ F_{12}^T & F_{22} \end{bmatrix}, \tag{2.5}$$

and pivots at that stage can only be chosen from within the fully summed block F_{11}. The problem occurs when it is impossible or numerically suicidal to eliminate all of F_{11} resulting in more work and storage (sometimes dramatically more) than forecast.

Our mixed pivoting approach is based on two phases. In the first phase, we perform numerical pivoting in the block F_{11} of fully summed variables until no remaining variables satisfy the numerical criterion. In the second phase, we eliminate the remaining fully summed variables, adding 1×1 perturbations defined using a threshold $\mu > 0$. Let us denote by a_{ii} the generic 1×1 pivot and by

$$P = \begin{bmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{bmatrix}$$

the generic 2×2 pivot.

Then, our pivoting strategy is defined as follows. We define:

$$g_1(i) = \frac{\max_{k\neq i} |a_{ki}|}{|a_{ii}|}, \tag{2.6}$$

and

$$g_2(i,j) = \left\| |P^{-1}| \begin{pmatrix} \max_{k\neq i,j} |a_{ki}| \\ \max_{k\neq i,j} |a_{kj}| \end{pmatrix} \right\|_\infty . \tag{2.7}$$

During the first phase of our algorithm, we perform the usual Duff-Reid algorithm. In this approach, a 1×1 pivot a_{ii} is considered to be stable if and only if

$$g_1(i) \leq 1/u, \tag{2.8}$$

and a 2×2 pivot is considered to be stable if and only if

$$g_2(i,j) \leq 1/u. \tag{2.9}$$

During the second phase we preselect a 1×1 and a 2×2 pivot and decide which will be eliminated. The choice between a 1×1 and a 2×2 pivot is done in three stages. Firstly, if we can eliminate a pivot and ensure a growth factor lower than $1 + 1/\mu$, where $\mu = \sqrt{\varepsilon}$, i.e., if $g_1(i) \leq 1/\mu$ or $g_2(i,j) \leq 1/\mu$, then we select the one with the lower growth factor. Secondly, if $g_1(i) \geq 1/\mu$ and $g_2(i,j) \geq 1/\mu$, (we cannot ensure a growth factor lower than $1 + 1/\mu$) then we compare the quantities $1/|a_{ii}|$ and $||P^{-1}||_\infty$ and we select the pivot associated with the smallest quantity. This second comparison is guided by the growth factor that would appear if we suppose that the largest off-diagonal entry is bounded by $\max_{ij} |a_{ij}|$. Finally, if no pivot can be chosen, we add some perturbation to make the factorization more stable. More precisely, a 1×1 pivot a_{ii} is selected and perturbed using $\delta = s(a_{ii})(\tau - |a_{ii}|)$ where s is the sign function and $\tau = \mu \max_{ij} |a_{ij}|$.

3 GMRES

The convergence of the GMRES method can be quite problematic for the general case even when we have a favourable distribution of the eigenvalues as in our case of the matrix $I - M^{-1}E$ (or $I - M^{-1}E$) (Arioli, Pták and Strakoš 1998).

However, because of the symmetry of A, the spectrum of $I - M^{-1}E$ (or $I - M^{-1}E$) has additional properties that make the behaviour more regular. Let us denote by $\widetilde{I}_m$ and $\widetilde{I}_{n-m}$ the $n \times m$ and $n \times (n-m)$ matrices corresponding to the first m columns and the last $n-m$ columns of the $n \times n$ identity matrix, respectively. We assume that the matrices E, M, and A are symmetrically permuted so that the nonzero entries of E are the first m diagonal entries of the matrix E.

We observe that $\widetilde{I}_{n-m}$ is a basis for $Ker(E)$.

Theorem 3.1. *Using the previous notation, if E is non-negative definite the spectrum of $I - M^{-1}E$ and $I - EM^{-1}$ are real.*

Proof. $M^{-1}Ex = \lambda x$ if and only if $Ex = \lambda Mx$. Since M is symmetric and nonsingular and E is positive semidefinite, λ is either zero or real, and (M, E) is an Hermitian definite pencil (see Z. Bai and J. Demmel and J. Dongarra and A. Ruhe and H. van der Vorst (2000), page 110). □

We can introduce the right preconditioned version of GMRES where we note that, although the vector z_k is computed, it is not stored.

Algorithm 3.1.

procedure [x] = right_Prec_gmres(A,M,b,maxit)
$x_0 = M^{-1}b$, $r_0 = b - Ax_0$ *and* $\beta = ||r_0||$
$v_1 = r_0/\beta$; $k = 0$; $r = r_0$
while $||r|| > \varepsilon\,(||b|| + ||A||\;||x_k||)$ *&* $k < maxit$
$k = k + 1$;
$z_k = M^{-1}v_k$; $w = Az_k$;
for $i = 1, \ldots, k$ *do*
$h_{i,k} = v_i^T w$;
$w = w - h_{i,k}v_i$;
end for;
$h_{k+1,k} = ||w||$;
$v_{k+1} = w/h_{k+1,k}$;
$V_k = [v_1, \ldots, v_k]$;
$H_k = \{h_{i,j}\}_{1 \le i \le j+1; 1 \le j \le k}$;
$y_k = \arg\min_y ||\beta e_1 - H_k y||$;
if $||\beta e_1 - H_k y_k|| \le \varepsilon\,(||b|| + ||A||\;||x_k||)$ *do*
$x_k = x_0 + M^{-1}V_k y_k$ *and* $r = b - Ax_k$;
end if
end while ;
end procedure.

4 Flexible GMRES

4.1 FGMRES vs GMRES

First, we introduce the right-preconditioned version of Flexible GMRES in Algorithm 4.1.
When $M_k = M\ \forall k$, the two algorithms 3.1 and 4.1 differ only in the computation of x_k where in Algorithm 4.1 the vectors z_k are stored in Z_k. Since Algorithm 4.1 needs both V_k and Z_k, it requires additional storage with respect to Algorithm 3.1. In Saad (2003) and Giraud, Gratton and Langou (2004), the convergence of FGMRES is analysed: if each matrix H_k is full rank, the algorithm converges to the solution.

Both algorithms are based on the following relations

$$C^{(k)} = [r_0, AZ_k] = V_{k+1} \begin{bmatrix} R_k \\ 0 \end{bmatrix}; \qquad V_j^T V_j = I_j \quad \forall j. \tag{4.10}$$

This corresponds to computing the orthogonal factorization of $C^{(k)}$ where each column $C_i^{(k)}$ is computed after the $(i-1)$-th step of the Gram-Schmidt orthogonalization process. Furthermore, we have

$$R_k = \begin{bmatrix} \beta e_1 & H_k \end{bmatrix} \tag{4.11}$$

$$\text{and} \qquad AZ_k = V_{k+1}H_k. \tag{4.12}$$

The vector y_k (in both Algorithms 3.1 and 4.1) is computed by a QR algorithm based on Givens rotations using the upper Hessenberg structure of H_k.

Algorithm 4.1.

procedure [x] =FGMRES(A,M_i,b,maxit)
$x_0 = M_0^{-1}b$, $r_0 = b - Ax_0$ *and* $\beta = ||r_0||$
$v_1 = r_0/\beta$; $k = 0$; $r = r_0$
while $||r|| > \varepsilon\,(||b|| + ||A||\;||x_k||)$ *&* $k < maxit$
$k = k + 1$;
$z_k = M_k^{-1}v_k$; $w = Az_k$;
for $i = 1, \ldots, k$ *do*
$h_{i,k} = v_i^T w$;
$w = w - h_{i,k}v_i$;
end for;
$h_{k+1,k} = ||w||$;
$v_{k+1} = w/h_{k+1,k}$;
$Z_k = [z_1, \ldots, z_k]$; $V_k = [v_1, \ldots, v_k]$;
$H_k = \{h_{i,j}\}_{1 \le i \le j+1; 1 \le j \le k}$;
$y_k = \arg\min_y ||\beta e_1 - H_k y||$;
if $||\beta e_1 - H_k y_k|| \le \varepsilon\,(||b|| + ||A||\;||x_k||)$ *do*
$x_k = x_0 + Z_k y_k$ *and* $r = b - Ax_k$;
end if
end while ;
end procedure.

5 Roundoff error analysis

In the following, we will denote by $c_p(n, j)$ functions that depend only on the dimension n and the integer j. We will avoid a precise formulation of these dependences, but we assume that each $c_p(n, j)$ grows moderately with n and j. Finally, if $B \in \mathbb{R}^{n\times m}, n \ge m$ is a full rank matrix, we denote by $\kappa(B) = ||B||\;||B^+||$ its spectral condition number where $B^+ = (B^T B)^{-1}B$.

The roundoff error analysis of both FGMRES and GMRES can be made in four stages:

1. Error analysis of the Arnoldi-Krylov process. We will analyse in detail the MGS approach, but similar results can be achieved using a Householder-based approach as presented by Walker (1988), Drkosova, Geenbaum, Rozložník and Strakoš (1995), and Arioli and Fassino (1996).

2. Error analysis of the Givens process used on the upper Hessenberg matrix H_k in order to reduce it to upper triangular form.

3. Error analysis of the computation of x_k in FGMRES and GMRES.

4. Use of the static pivoting properties and of (1.3) in order to have simpler expressions.

The first two stages of the roundoff error analysis are the same for both FGMRES and GMRES and the last two stages are specific to each algorithm.

5.1 FGMRES roundoff error analysis

In this section, we state the theorem presenting the main result that covers stages 1 to 3. The proofs are given in the appendices.

Theorem 5.1. *Applying Algorithm 4.1 to solve (1.1), using finite-precision arithmetic conforming to IEEE standard with relative precision ε and under the following four hypotheses:*

$$\sigma_{min}(\bar{H}_k) > c_1(k,1)\varepsilon\,||\bar{H}_k|| + \mathcal{O}(\varepsilon^{\,2}) \quad \forall k, \tag{5.13}$$

$$|\bar{s}_k| < 1-\varepsilon\,,\ \forall k, \tag{5.14}$$

where $\bar{s}_k$ are the sines computed during the Givens algorithm applied to $\bar{H}_k$ in order to compute $\bar{y}_k$, and

$$2.12(n+1)\varepsilon\ < 0.01 \qquad \textit{and} \qquad 18.53\varepsilon\, n^{\frac{3}{2}}\kappa(C^{(k)}) < 0.1\ \forall k \tag{5.15}$$

then there exists $\hat{k}$, $\hat{k} \le n$ such that, $\forall k \ge \hat{k}$, we have

$$||b - A\bar{x}_k|| \le c_2(n,k)\varepsilon\left(||b|| + ||A||\,||\bar{x}_0|| + ||A||\,||\bar{Z}_k||\,||\bar{y}_k||\right) + \mathcal{O}(\varepsilon^{\,2}). \tag{5.16}$$

Moreover, if $M_i = M, \forall i$, denoting by $\hat{W}_k = [M\bar{z}_1 - \bar{v}_1, \ldots, M\bar{z}_k - \bar{v}_k]$, and under the hypothesis

$$\rho = 1.3\,||\hat{W}_k|| + c_3(k,1)\varepsilon\,||M||\,||\bar{Z}_k|| < 1 \quad \forall k < \hat{k}, \tag{5.17}$$

we have:

$$\begin{aligned} ||b - A\bar{x}_k|| \ \le\ & c_4(n,k)\gamma\varepsilon\left(||b|| + ||A||\,||\bar{x}_0|| + \right. \\ & \left. ||A||\,||\bar{Z}_k||\left(||M(\bar{x}_k - \bar{x}_0)|| + \varepsilon\,||\bar{x}_0||\right)\right) + \mathcal{O}(\varepsilon^{\,2}) \end{aligned} \tag{5.18}$$

$$\gamma = \frac{1.3}{1-\rho}.$$

Proof. See Appendix A □

Remark 1. *We point out that if $\bar{x}_0 = 0$ and $M = I$, FGMRES is numerically equivalent to classical GMRES. Moreover, formula (5.18) implies that GMRES is normwise backward stable and we obtain the result of Paige, Rozložník and Strakoš (2006).*

5.1.1 Stage 4. Static pivoting case

We remark that the computation of the matrices $\hat{L}$ and $\hat{D}$ by `MA57` with static pivoting in floating-point arithmetic gives the following relations

$$\begin{cases} A + \delta A + \tau\Delta = M \\ ||\delta A|| \le c_5(n)\varepsilon\,||\,|\hat{L}|\,|\hat{D}|\,|\hat{L}^T|\,|| \\ ||\Delta|| \le 1. \end{cases} \tag{5.19}$$

Moreover, we take into account that

$$\begin{cases} (M + \delta M)\bar{x}_0 = b \qquad \text{where} \\ ||\delta M|| \le c_6(n)\varepsilon\ ||\,|\hat{L}|\,|\hat{D}|\,|\hat{L}^T|\,||. \end{cases} \tag{5.20}$$

Relations (5.19) and (5.20) follow by the use of standard techniques similar to those for the roundoff error analysis of Gaussian elimination (Wilkinson 1965, Golub and Van Loan 1989, Higham 2002) . We point out that, from (5.19), the perturbation δA must have a norm smaller than τ in order not to dominate the global error that would make the diagonal perturbation ineffective. Therefore, it is reasonable to assume that

$$\max\left(c_5(n), c_6(n)\right)\varepsilon\,||\,|\hat{L}|\,|\hat{D}|\,|\hat{L}^T|\,|| < \tau \tag{5.21}$$

Theorem 5.2. *Under the hypotheses of Theorem 5.1, hypothesis (5.21) and under the hypotheses*

$$c_7(n,k)\gamma\varepsilon\,||A||\,||\bar{Z}_k|| < 1 \quad \forall k < \hat{k} \tag{5.22}$$

and

$$\max\{||M^{-1}||, ||\bar{Z}_k||\} \leq \frac{\tilde{c}}{\tau} \tag{5.23}$$

where $\tilde{c}$ is a constant, we have

$$||b - A\bar{x}_k|| \leq 2\mu\varepsilon\,(||b|| + ||A||\,(||\bar{x}_0|| + ||\bar{x}_k||)) + \mathcal{O}(\varepsilon^{\,2}). \tag{5.24}$$

$$\mu = \frac{c_7(n,k)\gamma}{1 - c_7(n,k)\varepsilon\,\gamma||A||\,||\bar{Z}_k||}.$$

Proof. See Appendix B □

Remark 2. *We point out that hypothesis (5.23), as we will see in Section 6, is satisfied for a moderate value of $\tilde{c}$ by our static pivoting strategy on all our test problems. However, the upper bound (5.18) is more general and also gives a good estimate of the error when $\tau << \sqrt{\varepsilon}$. In all our experiments, we have $||\bar{Z}_k||\,||M(\bar{x}_k - \bar{x}_0)|| < 10^4$.*

5.2 GMRES roundoff error analysis

For the sake of simplicity, we make the assumption that the solution of the linear system

$$Mq = fl(\bar{V}_k\bar{y}_k),$$

is followed by a few steps of iterative refinement in order to guarantee a good local backward stability. Under this assumption, we have

Theorem 5.3. *Applying Algorithm 3.1 to solve (1.1), using finite-precision arithmetic conforming to IEEE standard with relative precision ε and under the hypotheses (5.13), (5.14), and (5.15) of Theorem 5.1 and if*

$$c_8(n,1)\varepsilon\ \kappa(M) < 1 \tag{5.25}$$

then there exists $\hat{k}$, $\hat{k} \leq n$ such that, $\forall k \geq \hat{k}$, we have

$$\begin{aligned} ||b - A\bar{x}_k|| \leq\ & c_9(n,k)\chi\varepsilon \Big\{ ||b|| + ||A||\,||\bar{x}_0|| + ||AM^{-1}||\,||M||\,||\bar{x}_k - \bar{x}_0|| + \\ & \Big[||A||\,||\bar{Z}_k|| + ||AM^{-1}||\,||\,|\hat{L}|\,|\hat{D}|\,|\hat{L}^T|\,|| \Big] \times \\ & \Big[||M(\bar{x}_k - \bar{x}_0)|| + n\varepsilon\ ||M||\,(||\bar{x}_k - \bar{x}_0|| + ||\bar{x}_0||) \Big] \Big\} + \mathcal{O}(\varepsilon^{\,2}) \end{aligned} \tag{5.26}$$

$$\chi = \frac{1.3}{1 - 1.3k\varepsilon}.$$

Proof. See Appendix C □

Remark 3. *Both (5.18) and (5.26) do not depend on the special initial choice $\bar{x}_0 = M^{-1}b$. They will be true for any choice of $\bar{x}_0$. In both FGMRES and GMRES, a restart with a new $\bar{x}_0 = \bar{x}_k$ can improve the situation reducing the upper bounds significantly.*

Remark 4. *Finally, we point out that substituting the values of $c_4(n,k)\gamma$ and $c_9(n,k)\chi$ in (5.18) and (5.26) with $\hat{c} = \max(c_4(n,k)\gamma, c_9(n,k)\chi)$, (5.26) becomes an upper bound for (5.18). The numerical experiments in Section 6 support this observation.*

5.2.1 Stage 4. Static pivoting case

From (5.19) and (5.21), taking into account that $\tau = \varepsilon^{\sigma}||A||$ with $0 < \sigma \leq 1$, we have

$$||M|| \leq ||A|| + \tau + c_{10}(n)\,\varepsilon\;||\,|\hat{L}|\,|\hat{D}|\,|\hat{L}^T|\,|| \leq ||A|| + 2\,\tau \leq 3\,||A||. \tag{5.27}$$

From (5.26), we have

$$\begin{aligned} ||b - A\bar{x}_k|| \leq \;& c_{11}(n,k)\,\chi\,\varepsilon\,\Big[\;||b|| + ||A||\,||\bar{x}_0|| + ||AM^{-1}||\,||A||\,||\bar{x}_k - \bar{x}_0|| + \\ & ||A||\,||\bar{Z}_k||\,||\,M(\bar{x}_k - \bar{x}_0)|| + \\ & ||AM^{-1}||\,||\,|\hat{L}|\,|\hat{D}|\,|\hat{L}^T|\,||\,||A||\,||\bar{x}_k - \bar{x}_0||\;\Big] + \mathcal{O}(\varepsilon^{\,2}). \end{aligned} \tag{5.28}$$

Moreover, we have

$$AM^{-1} = I - \tau\Delta M^{-1} - c_5(n)\,\varepsilon\;||\,|\hat{L}|\,|\hat{D}|\,|\hat{L}^T|\,||$$

and, from (5.23) and assuming (5.21) is satisfied ,

$$||AM^{-1}|| \leq 1 + \tilde{c}.$$

Thus we have from (5.28)

$$\begin{aligned} ||b - A\bar{x}_k|| \leq \;& c_{12}(n,k)\,\chi\,\varepsilon\,\Big[\;||b|| + ||A||\,||\bar{x}_0|| + ||A||\,||\bar{Z}_k||\,||\,M(\bar{x}_k - \bar{x}_0)|| + \\ & \Big[\;1 + ||\,|\hat{L}|\,|\hat{D}|\,|\hat{L}^T|\,||\;\Big]\,||A||\,||\bar{x}_k - \bar{x}_0||\;\Big] + \mathcal{O}(\varepsilon^{\,2}). \end{aligned} \tag{5.29}$$

$$\chi = \frac{1.3}{1 - 1.3k\varepsilon}$$

Remark 5. *Formula (5.29) suggests that we might expect GMRES to have a final residual larger than the final residual in FGMRES. In particular, we could expect that a factor proportional to* $\varepsilon\;||A||\,||\,|\hat{L}|\,|\hat{D}|\,|\hat{L}^T|\,||\,||\bar{x}_k - \bar{x}_0||$ *will appear.*

6 Numerical experiments

6.1 Test problems

We have run FGMRES and GMRES on several test problems. For the sake of simplicity, we only present the results obtained on particularly tough problems where iterative refinement fails for several different static pivoting strategies and values of τ (Duff and Pralet 2005). We decided to omit the other results either because they are very similar to our results or because iterative refinement produced a scaled residual that is at rounding error level.

The dimension, the number of nonzero entries and the origin of the three test problems that we use to illustrate the theory presented in Section 5 are shown in Table 6.1.

6.2 Numerical results

The smallest example TUMA1 presents some interesting features and illustrates the behaviour of test examples where iterative refinement converges for the range of τ we are interested in. In Figure 6.1, we summarise the behaviour of $||\;|L|\,|D|\,|L^T|\;||$ vs $1/\tau$ for all our test examples. We point out that $||\;|L|\,|D|\,|L^T|\;||$ can grow much more than $1/\tau$ because of the threshold pivoting strategy, and this growth is the reason for the poor behaviour of GMRES.

	n	nnz	$\|\|A\|\|_\infty$	$\|\|A\|\|$	Description
CONT201	80595	239596	8.2	8.1	KKT matrix Convex QP (M2)
CONT300	180895	562496	8.2	8.1	KKT matrix Convex QP (M2)
TUMA1	22967	76199	7.8	4.9	Mixed-Hybrid finite-element

Table 6.1: Test problems

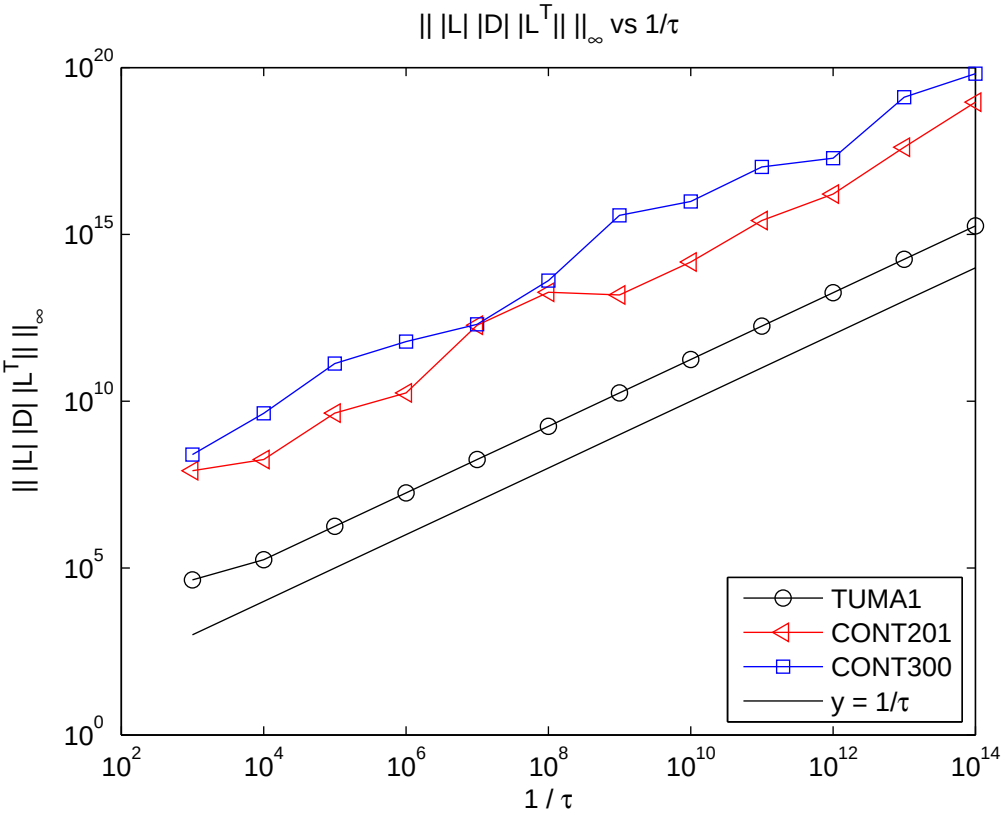

Figure 6.1: $||\,|L|\,|D|\,|L^T|\,||$ vs $1/\tau$

In Tables 6.2, 6.3, and 6.4, we present the parameters and the scaled residual

$$\frac{||b - A\bar{x}_k||}{||b|| + ||A||_\infty\, ||\bar{x}_k||} \tag{6.30}$$

for right-preconditioned GMRES (GMRES column) and Flexible GMRES (FGMRES column), for the test problems TUMA1, CONT201, and CONT300 respectively. We also show the values of $||\bar{Z}_k||$ and $||M(\bar{x}_k - \bar{x}_0)||$ at convergence, as they are crucial quantities for the residual bounds in Theorems 5.1 and 5.3.

In all test examples, the values of $||\bar{Z}_k||$ and $||M(\bar{x}_k - \bar{x}_0)||$ for FGMRES show that

$$||\bar{Z}_k||\,||M(\bar{x}_k - \bar{x}_0)|| < 10^4$$

	$\frac{\|\|b - A\bar{x}_k\|\|}{\|\|b\|\| + \|\|A\|\|\|\bar{x}_k\|\|}$			$\|\|M(\bar{x}_k - \bar{x}_0)\|\|$		
τ	GMRES (#It)	FGMRES (#It)	$\|\|Z_k\|\|$	FGMRES	GMRES	$\|\|\|L\|\|D\|\|L^T\|\|\|_\infty$
1.0e-03	1.0e-14 (26)	7.2e-17 (9)	1.2e+02	3.5e-03	3.5e-03	4.4e+04
1.0e-04	1.8e-16 (6)	3.1e-17 (6)	4.7e+01	4.4e-04	4.4e-04	1.8e+05
1.0e-05	1.3e-16 (5)	1.9e-17 (5)	4.4e+01	4.5e-05	4.5e-05	1.8e+06
1.0e-06	1.3e-16 (4)	1.9e-17 (4)	4.4e+01	4.5e-06	4.5e-06	1.8e+07
1.0e-07	1.2e-16 (3)	2.0e-17 (3)	4.3e+01	4.5e-07	4.5e-07	1.8e+08
1.0e-08	1.3e-16 (3)	1.8e-17 (3)	4.3e+01	4.5e-08	4.5e-08	1.8e+09
1.0e-09	2.8e-15 (31)	1.8e-17 (3)	2.6e+01	4.0e-08	4.0e-08	1.8e+10
1.0e-10	4.2e-13 (31)	1.8e-17 (3)	8.8e+00	4.0e-07	4.0e-07	1.8e+11
1.0e-11	1.0e-10 (31)	6.2e-17 (3)	6.8e+00	4.0e-06	4.0e-06	1.8e+12
1.0e-12	1.0e-08 (31)	2.2e-17 (4)	3.2e+01	4.3e-05	4.3e-05	1.8e+13
1.0e-13	2.4e-07 (31)	1.9e-17 (6)	1.3e+02	3.9e-04	3.9e-04	1.8e+14
1.0e-14	8.6e-06 (31)	2.1e-17 (10)	1.8e+02	4.3e-03	4.3e-03	1.8e+15

Table 6.2: TUMA_1 results

when $\tau = 10^{-9}$ (see Table 6.4). However, for the remaining values of τ the latter product is less than 2×10^2.

In Figures 6.2 and 6.3, we show the convergence histories of the scaled residual (6.30) for FGMRES and the right-preconditioned version of GMRES for the CONT201 and CONT300 test problems.

Firstly, we point out that right-preconditioned GMRES is not backward stable. The FGMRES dependence on τ shows that the best tradeoff is obtained for the recommended choice of $\tau = 10^{-8}$ where both iterative refinement and right-preconditioned GMRES are not able to obtain a scaled residual close to ε.

Both FGMRES and right-preconditioned GMRES can benefit from a restarting process. In particular, the restarted right-preconditioned GMRES converges with a scaled residual at machine precision for a small value of the restart parameter. In Figures 6.4 and 6.5, we compare the restarted right-preconditioned GMRES with FGMRES for $\tau = 10^{-8}$ for the CONT201 and CONT300 test examples respectively. In the same figures, we show the behaviour of classical iterative refinement which is unable to obtain a scaled residual close to rounding. Furthermore, we show in Figure 6.6 that, even with restarting, the preconditioned GMRES approach is not robust.

7 Conclusions

We have shown by experiment and analysis that FGMRES is a powerful method for obtaining the solution of sets of sparse linear equations when a direct method using static pivoting has been used to factorize the matrix. This factorization is then used as a preconditioner for an iterative method. In particular, we show that GMRES converges when Richardson's method (iterative

	$\frac{\|\|b - A\bar{x}_k\|\|}{\|\|b\|\| + \|\|A\|\|\|\bar{x}_k\|\|}$			$\|\|M(\bar{x}_k - \bar{x}_0)\|\|$		
τ	GMRES (#It)	FGMRES (#It)	$\|\|Z_k\|\|$	FGMRES	GMRES	$\|\|\|L\|\|D\|\|L^T\|\|\|_\infty$
1.0e-03	1.8e-05 (31)	9.8e-06 (31)	*	1.5e-04	7.1e-04	8.3e+07
1.0e-04	2.0e-07 (31)	2.0e-07 (31)	*	1.9e-05	1.5e-05	1.8e+08
1.0e-05	1.8e-12 (31)	1.1e-16 (30)	4.1e+05	1.3e-05	5.9e-06	4.4e+09
1.0e-06	1.1e-11 (31)	2.1e-16 (15)	2.7e+06	7.8e-07	7.8e-07	1.8e+10
1.0e-07	4.8e-11 (31)	1.8e-16 (9)	1.4e+08	8.7e-08	8.7e-08	1.9e+12
1.0e-08	2.7e-10 (31)	5.8e-17 (6)	2.1e+07	1.3e-06	1.3e-06	1.8e+13
1.0e-09	1.8e-09 (31)	4.5e-17 (5)	1.1e+07	1.3e-06	1.3e-06	1.5e+13
1.0e-10	3.2e-09 (31)	7.2e-17 (4)	3.4e+05	9.2e-06	9.2e-06	1.5e+14
1.0e-11	2.1e-09 (31)	4.5e-17 (4)	1.9e+03	2.8e-04	2.8e-04	2.6e+15
1.0e-12	4.5e-07 (31)	3.8e-17 (5)	2.0e+02	9.5e-04	9.5e-04	1.6e+16
1.0e-13	1.3e-04 (31)	2.6e-16 (8)	1.6e+02	1.1e-02	1.1e-02	4.1e+17
1.0e-14	2.3e-01 (31)	2.5e-14 (12)	4.3e+02	1.0e-02	1.9e-02	9.2e+18

Table 6.3: CONT_201 results

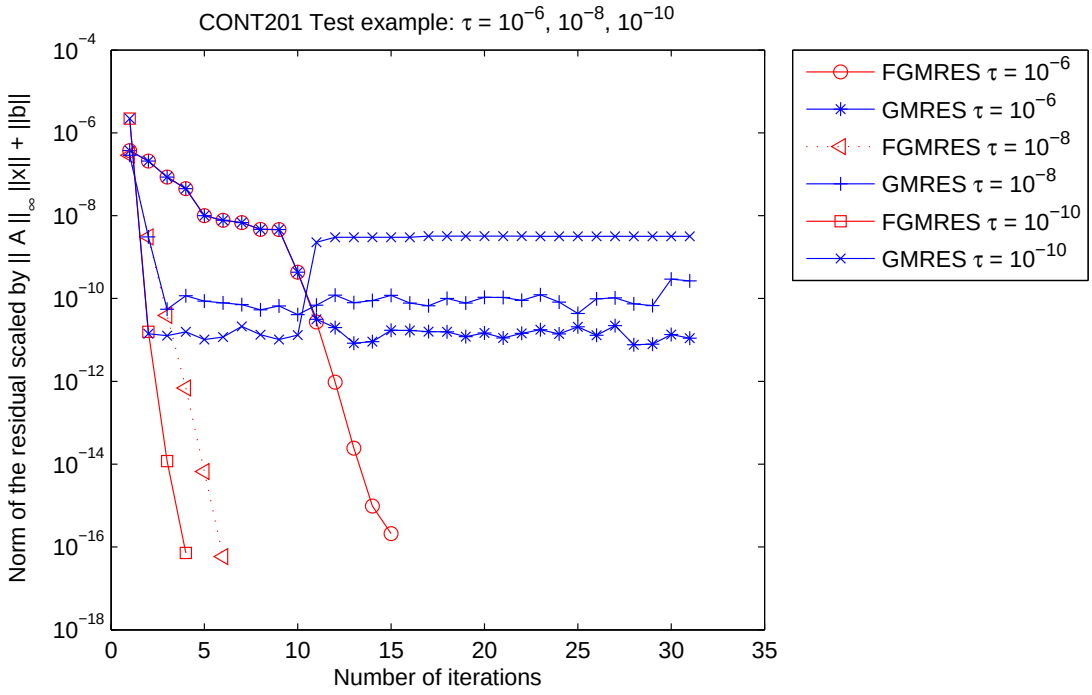

Figure 6.2: GMRES vs FGMRES on CONT201 test example: $\tau = 10^{-6}, 10^{-8}, 10^{-10}$

refinement) does not, that restarted GMRES performs better, and that FMGRES succeeds when these other methods fail. We also see that the convergence of FGMRES is much less sensitive to

	$\frac{\|\|b - A\bar{x}_k\|\|}{\|\|b\|\| + \|\|A\|\|\|\bar{x}_k\|\|}$			$\|\|M(\bar{x}_k - \bar{x}_0)\|\|$		
τ	GMRES (#It)	FGMRES (#It)	$\|\|Z_k\|\|$	FGMRES	GMRES	$\|\|\|L\|\|D\|\|L^T\|\|\|_\infty$
1.0e-03	3.6e-05 (31)	2.5e-05 (31)	*	1.3e-04	8.7e-04	2.5e+08
1.0e-04	5.5e-07 (31)	5.5e-07 (31)	*	2.8e-05	6.5e-05	4.3e+09
1.0e-05	8.7e-09 (31)	8.7e-09 (31)	*	6.1e-06	3.7e-06	1.4e+11
1.0e-06	6.9e-11 (31)	1.4e-16 (23)	3.0e+06	9.8e-07	5.7e-07	6.2e+11
1.0e-07	2.1e-10 (31)	8.2e-17 (12)	7.6e+06	2.3e-07	2.3e-07	2.0e+12
1.0e-08	1.4e-08 (31)	1.2e-16 (8)	7.5e+07	1.8e-06	1.8e-06	4.1e+13
1.0e-09	1.6e-05 (31)	8.8e-17 (8)	3.7e+07	2.8e-04	2.8e-04	3.7e+15
1.0e-10	6.8e-07 (31)	4.1e-17 (6)	3.8e+05	3.6e-04	3.6e-04	9.6e+15
1.0e-11	1.6e-06 (31)	8.7e-17 (5)	1.4e+03	5.3e-03	5.3e-03	1.0e+17
1.0e-12	1.1e-06 (31)	2.7e-16 (5)	1.5e+02	1.0e-02	1.0e-02	1.9e+17
1.0e-13	3.4e-03 (31)	9.2e-16 (7)	1.3e+02	1.9e-01	1.9e-01	1.3e+19
1.0e-14	1.4e-01 (31)	1.8e-14 (12)	2.1e+02	4.7e-02	4.7e-02	6.6e+19

Table 6.4: CONT_300 results

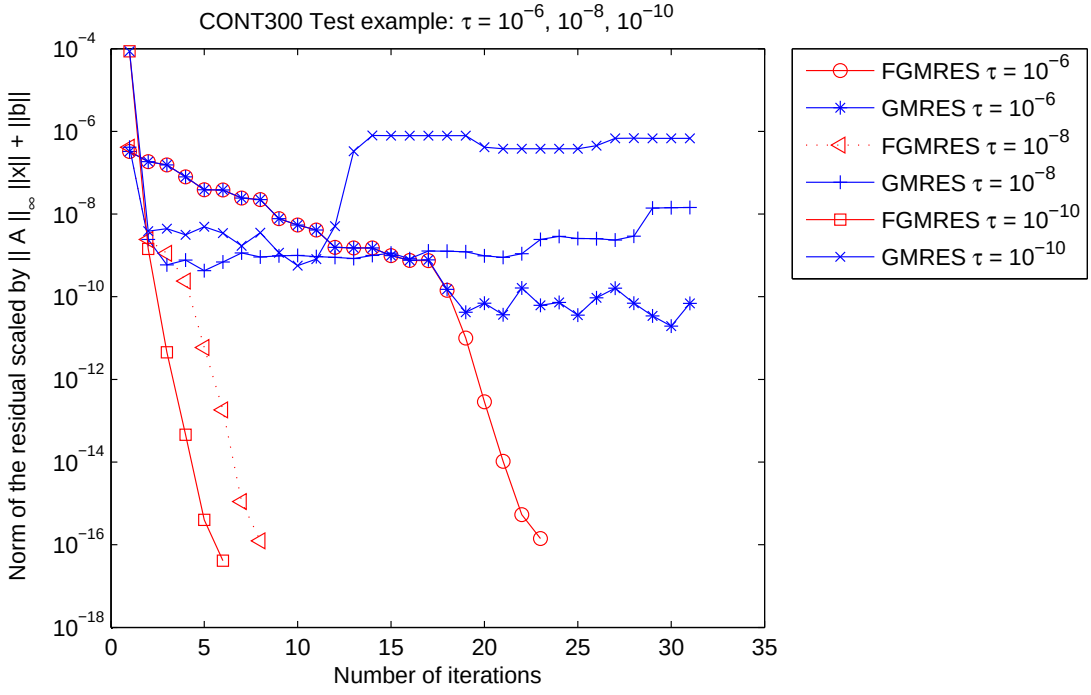

Figure 6.3: GMRES vs FGMRES on CONT300 test example: $\tau = 10^{-6}, 10^{-8}, 10^{-10}$

the choice of the static pivoting parameter.

Our analysis gives sufficient conditions for convergence and we observe that often FGMRES

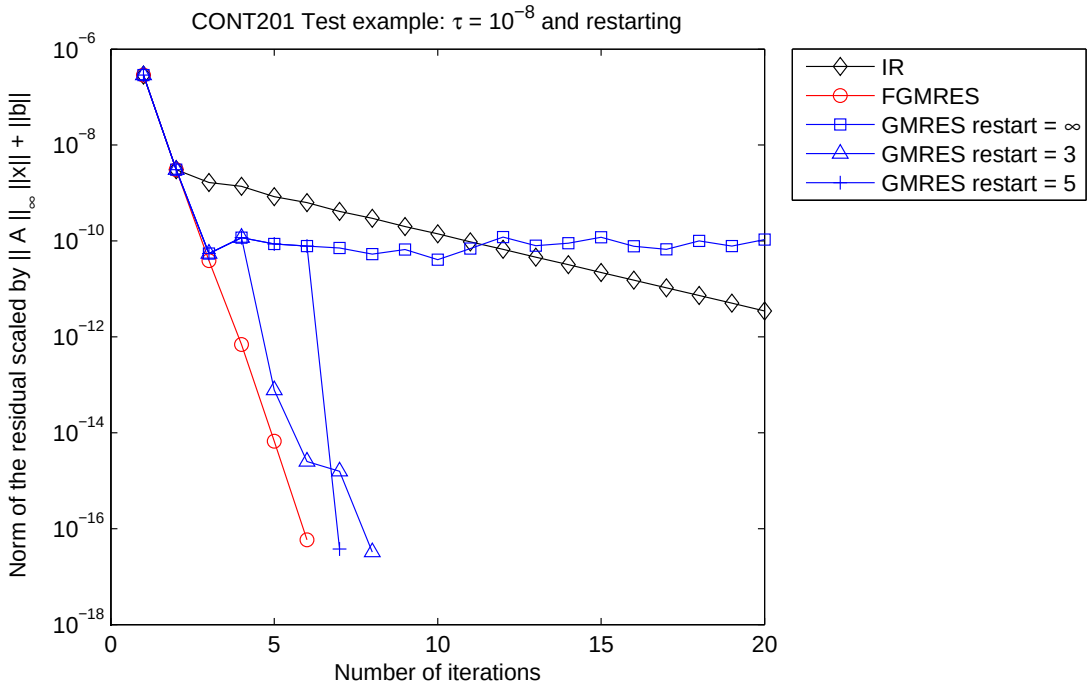

Figure 6.4: Restarted GMRES vs FGMRES on CONT201 test example: $\tau = 10^{-8}$

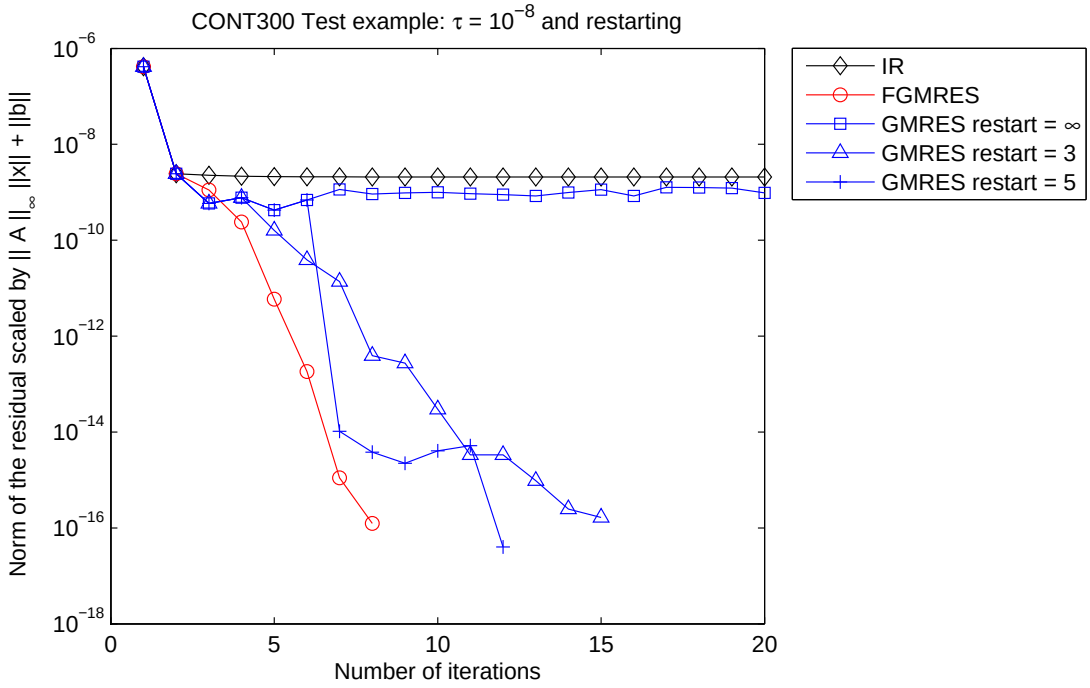

Figure 6.5: Restarted GMRES vs FGMRES on CONT300 test example: $\tau = 10^{-8}$

converges even when our hypotheses are not satisfied. The main reason for this is that the bound $|| \, |L| \, |D| \, |L^T| \, ||$ on the norm of the solution of the preconditioned system is rather crude. Indeed, taking into account that $||M|| \approx ||A||$, we see that there is severe cancellation in the product and

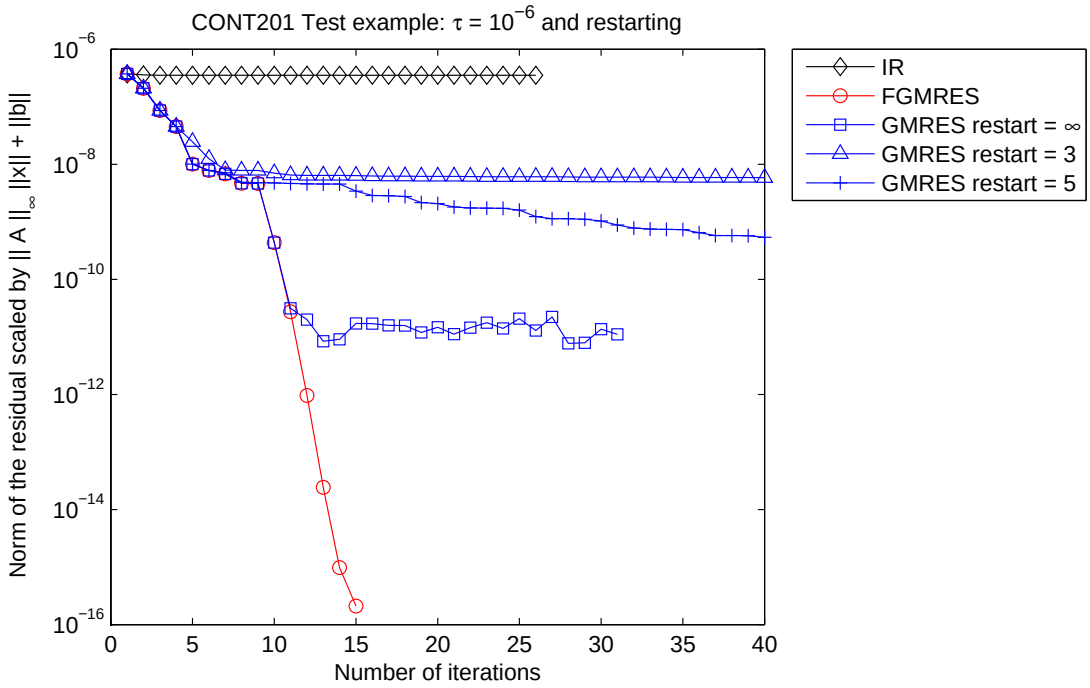

Figure 6.6: Restarted GMRES on CONT201 test example: $\tau = 10^{-6}$

that small entries in D are usually balanced by large entries in L.

However, we see clearly that the analysis showing that stronger hypotheses are required for the convergence of preconditioned GMRES is reflected in the poorer convergence of this method, even when restarting is used.

A great benefit of our analysis and experiments is that the FGMRES iterative method can be used with confidence to solve systems preconditioned with a static pivoting factorization and that the method is far less sensitive to the choice of static pivoting parameter than either GMRES or iterative refinement.

Acknowledgment We would like to thank the anonymous referees for their helpful comments that have hopefully made the paper easier to read.

References

Arioli, M. and Fassino, C. (1996), 'Roundoff Error Analysis of Algorithms Based on Krylov Subspace Methods', *BIT* **36**, 189–206.

Arioli, M., Pták, V. and Strakoš, Z. (1998), 'Krylov sequences of maximal length and convergence of GMRES', *BIT* **38**, 1–9.

Z. Bai and J. Demmel and J. Dongarra and A. Ruhe and H. van der Vorst (2000), *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Björck, Å. and Paige, C. C. (1992), 'Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm', *SIAM J. Matrix Anal. Appl.* **13**(1), 176–190.

Demmel, J., Hida, Y., Kahan, W., Li, X., Mukherjee, S. and Riedy, E. (2005), Error bounds from extra precise iterative refinement, ACM Trans. Math. Softw., 32 (2006), pp. 325–351.

Drkosova, J., Geenbaum, A., Rozložník, M., and Strakoš, Z. (1995), 'Numerical stability of GMRES method', *BIT* **35**, 308–330.

Duff, I. S. (2004),'MA57 – A code for the solution of symmetric indefinite systems', *TOMS* **30**, 118–144.

Duff, I. S. and Pralet, S. (2005), Towards a stable static pivoting strategy for the sequential and parallel solution of sparse symmetric indefinite systems, Technical Report TR/PA/05/26, CERFACS, Toulouse, France. Also available as RAL Report RAL-TR-2005-007 and IRIT Report RT/TLSE/05/04.

Giraud, L. and Langou, J. (2002), 'When modified Gram-Schmidt generates a well-conditioned set of vectors', *IMA J. Numer. Anal.* **22**, 521–528.

Giraud, L., Gratton, S. and Langou, J. (2004), A note on relaxed and flexible GMRES, Technical Report TR/PA/04/41, CERFACS, Toulouse, France.

Golub, G. H. (1965), 'Numerical methods for solving linear least squares problems', *Numer. Math.* **7**, 206–216.

Golub, G. H. and Van Loan, C. F. (1989), *Matrix Computations*, second edn, Johns Hopkins University Press, Baltimore, MD, USA.

Higham, N. J. (2002), *Accuracy and Stability of Numerical Algorithms, Second Edition*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Paige, C., Rozložník, M. and Strakoš, Z. (2006), 'Modified Gram-Schmidt (MGS), least squares, and backward stability of MGS-GMRES', *SIAM Journal on Matrix Analysis and Applications* **28**(1), 264–284.

Saad, Y. and Schultz M. H. (1986), 'GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems' *SIAM J. Sci. Stat. Comput.* **7**, 856–869.

Saad, Y. (1993), 'A flexible inner-outer preconditioned GMRES algorithm' *SIAM J. Sci. Stat. Comput.* **14**, 461–469.

Saad, Y. (2003), *Iterative Methods for Sparse Linear Systems Second Edition*, Society for Industrial and Applied Mathematics.

Skeel, R. D. (1980), 'Iterative refinement implies numerical stability for Gaussian elimination', *Math. Comp.* **35**(151), 817–832.

Walker, H. F. (1988), 'Implementation of the GMRES method using Householder transformations', *SIAM J. Sci. Stat. Comput.* **9**(1), 152–163.

Wilkinson, J. H. (1965), *The Algebraic Eigenvalue Problem*, Oxford University Press.

A : Proof of Theorem 5.1

Stage 1

By standard techniques (Higham 2002), the computed matrix by vector products satisfy the relations

$$\begin{aligned} fl(r_0) &= r_0 + f_1 \qquad ||f_1|| \leq c_{13}(n,1)\varepsilon\,(||b|| + ||A||\,||\bar{x}_0||) + \mathcal{O}(\varepsilon^{\,2}), \qquad \text{(A.31)} \\ fl(A\bar{Z}_k) &= A\bar{Z}_k + F_k, \qquad ||F_k|| \leq c_{14}(n,k)\varepsilon\,||A||\,||\bar{Z}_k|| + \mathcal{O}(\varepsilon^{\,2}), \qquad \text{(A.32)} \end{aligned}$$

where $\bar{x}_0 = fl(x_0)$.

Following Björck and Paige (1992) and Giraud and Langou (2002), the Gram-Schmidt orthogonalization process applied to $fl(C^{(k)})$ computes an upper triangular matrix $\overline{R}_k$ for which there exists an orthonormal matrix $\hat{V}_{k+1}$ that satisfies the relations:

$$\left\{ \begin{array}{l} [fl(r_0); fl(A\bar{Z})] + [f_2; E_k] = \hat{V}_{k+1}\overline{R}_k, \qquad \hat{V}_{k+1}^T\hat{V}_{k+1} = I_{k+1} \\ ||f_2|| \leq c_{15}(n,1)\varepsilon\,||r_0|| + \mathcal{O}(\varepsilon^{\,2}) \qquad ||E_k|| \leq c_{16}(n,k)\varepsilon\,||A\bar{Z}_k|| + \mathcal{O}(\varepsilon^{\,2}) \end{array} \right. \qquad \text{(A.33)}$$

under the hypothesis (5.15).

By combining (A.31), (A.32), and (A.33), we have

$$\left\{ \begin{array}{l} [r_0; A\bar{Z}_k] + [f_1 + f_2; F_k + E_k] = \hat{V}_{k+1}\overline{R}_k, \\ ||f_1 + f_2|| \leq c_{17}(n,1)\varepsilon\,(||b|| + ||A||\,||\bar{x}_0||) + \mathcal{O}(\varepsilon^{\,2}) \quad \text{and} \\ ||F_k + E_k|| \leq c_{18}(n,k)\varepsilon\,(||A\bar{Z}_k|| + ||A||\,||\bar{Z}_k||) + \mathcal{O}(\varepsilon^{\,2}). \end{array} \right. \qquad \text{(A.34)}$$

Stage 2

The second part of FGMRES computes the vector $\bar{y}_k = fl(y_k)$ by Givens or Householder algorithms (Golub 1965, Golub and Van Loan 1989). The vector $\bar{y}_k$ satisfies the following relations

$$\left\{ \begin{array}{l} \bar{y}_k = \arg\min_y ||\bar{\beta}e_1 + g^{[k]} - (\bar{H}_k + \Delta\bar{H}_k)y||, \\ ||\Delta\bar{H}_k|| \leq c_{19}(k,1)\varepsilon\,||\bar{H}_k|| + \mathcal{O}(\varepsilon^{\,2}) \text{ and } ||g^{[k]}|| \leq c_{20}(k,1)\varepsilon\,\bar{\beta} + \mathcal{O}(\varepsilon^{\,2}), \end{array} \right. \qquad \text{(A.35)}$$

where $\bar{\beta} = (\overline{R}_k)_{11} = ||r_0 + f_1 + f_2||$ and the columns of $\bar{H}_k$ are columns $2, \ldots, k+1$ of $\overline{R}_k$. Note that $\bar{\beta}$ is independent of k because at step k we only compute the kth column of $\overline{R}_k$, leaving the earlier columns unchanged. The computation of $\bar{y}_k$ is performed in two stages and contributions to the matrix $\Delta\bar{H}_k$ come from two sources.

First, a sequence of Givens (or Householder) rotations $G^{(i)}$ is computed in order to reduce the matrix $\bar{H}$ to the upper triangular form U_k. The floating-point computation of the matrices $G^{(i)}$ gives

$$fl(G^{(i)}) = \bar{G}^{(i)} = \begin{bmatrix} I_{i-1} & & & \\ & \bar{c}_i & -\bar{s}_i & \\ & \bar{s}_i & \bar{c}_i & \\ & & & I_{n-i-1} \end{bmatrix} \quad i = 1, \ldots, k$$

$$\bar{c}_i = fl\left(\frac{(\bar{H}_k)_{i,i}}{\sqrt{(\bar{H}_k)_{i,i}^2 + (\bar{H}_k)_{i+1,i}^2}}\right) \qquad \text{and} \qquad \bar{s}_i = fl\left(\frac{(\bar{H}_k)_{i+1,i}}{\sqrt{(\bar{H}_k)_{i,i}^2 + (\bar{H}_k)_{i+1,i}^2}}\right)$$

The $\bar{G}^{(i)}$ matrices are also applied to the vector $\bar{\beta}e_1$ and, from the error analysis presented by Wilkinson (1965), the floating-point arithmetic will produce an exact orthogonal matrix $G^{[k]}$ such that

$$\begin{aligned} fl(\bar{G}^{(k)} \cdots \bar{G}^{(1)} \bar{\beta} e_1) &= G^{[k]}(\bar{\beta}e_1 + g^{[k]}) \\ fl(\bar{G}^{(k)} \cdots \bar{G}^{(1)} \bar{H}_k) &= G^{[k]}(\bar{H}_k + \Delta \bar{H}_k^{(1)}), \\ ||\Delta \bar{H}_k^{(1)}|| &\leq c_{21}(k,1)\varepsilon\, ||\bar{H}_k|| + \mathcal{O}(\varepsilon^{\,2}). \end{aligned}$$

Second, the $\bar{y}_k$ vector is computed by solving the upper triangular system. The standard backward substitution algorithm will introduce an additional perturbation $\Delta \bar{H}_k^{(2)}$ of $\bar{H}_k$ but will leave the perturbation $g^{[k]}$ untouched. The perturbation $\Delta \bar{H}_k^{(2)}$ will also have the same upper Hessenberg structure of $\bar{H}_k$ and

$$||\Delta \bar{H}_k^{(2)}|| \leq c_{22}(k,1)\varepsilon\, ||\bar{H}_k|| + \mathcal{O}(\varepsilon^{\,2}).$$

This follows from the upper triangular structure of the perturbation to U_k induced by the backward substitution algorithm, and from the structure and orthogonality of $G^{[k]}$. Finally, we point out that in the relations (A.35) we have see (5.13)

$$\begin{cases} \Delta \bar{H}_k = \Delta \bar{H}_k^{(1)} + \Delta \bar{H}_k^{(2)}, & \text{and} \\ c_1(k,1) = c_{21}(k,1) + c_{22}(k,1). \end{cases}$$

where $c_1(k,1)$ is the constant of (5.13) Moreover, because of the special structure of $\bar{\beta}e_1$ and the orthogonality of $G^{[k]}$ we have

$$g_j^{[k]} = 0 \qquad j = k+1, \ldots, n,$$

and, denoting by $\bar{h}^{[k]} = fl(\bar{G}^{(k)} \cdots \bar{G}^{(1)} \bar{\beta} e_1)$, we have

$$\begin{aligned} \alpha_k &= ||G^{[k]}(\bar{\beta}e_1 + g^{[k]} - (\bar{H}_k + \Delta \bar{H}_k)\bar{y}_k)|| \\ &= ||(G^{[k]}(\bar{\beta}e_1 + g^{[k]}))_{k+1}|| = |\bar{h}_{k+1}^{[k]}|. \end{aligned} \tag{A.36}$$

A direct analysis of $\bar{h}^{[k]}$ shows that

$$\begin{cases} \bar{h}_1^{[1]} &= \bar{\beta}\bar{c}_1(1+\mu_1) & |\mu_1| \leq \varepsilon \\ \bar{h}_2^{[1]} &= \bar{\beta}\bar{s}_1(1+\rho_1) & |\rho_1| \leq \varepsilon \\ \bar{h}_j^{[k]} &= \bar{h}_j^{[k-1]} & j = 1, \ldots, k-1, \quad k \geq 2, \\ \bar{h}_k^{[k]} &= \bar{h}_k^{[k-1]}\bar{c}_k(1+\mu_k) & |\mu_k| \leq \varepsilon \\ \bar{h}_{k+1}^{[k]} &= \bar{h}_k^{[k-1]}\bar{s}_k(1+\rho_k) & |\rho_k| \leq \varepsilon\,. \end{cases}$$

Therefore, from using (A.33), (A.34), (A.35), (A.36), and the orthogonality of

$$\widetilde{V}_{k+1} = \hat{V}_{k+1} G^{[k]T},$$

we have

$$\begin{aligned} \alpha_k &= ||\widetilde{V}_{k+1} G^{[k]}(\bar{\beta}e_1 + g^{[k]} - (\bar{H}_k + \Delta \bar{H}_k)\bar{y}_k)|| \\ &= ||r_0 + f_1 + f_2 + \hat{V}_{k+1} g^{[k]} - A(\bar{Z}_k + A^{-1}(F_k + E_k + \hat{V}_{k+1}\Delta \bar{H}_k))\bar{y}_k|| \\ &= ||r_0 + \delta r_0 - A\big(\bar{Z}_k + \hat{Z}_k\big)\bar{y}_k||, \end{aligned}$$

where $\delta r_0 = f_1 + f_2 + f_3$ and $\hat{Z}_k = A^{-1}(F_k + E_k + \hat{V}_k \Delta \bar{H}_k)$. Then

$$\alpha_k = \alpha_{k-1} |\bar{s}_k| (1 + \rho_k), \quad |\rho_k| \leq \varepsilon \,. \tag{A.37}$$

Thus, under the hypotheses (5.14), (5.13), and (5.15) and from (A.35), the matrices $\left(\bar{Z}_k + \hat{Z}_k\right)$ have full rank for all k, i.e.

$$\bar{H}_k + \Delta \bar{H}_k$$

is full rank $\forall k$, and the values of α_k converge monotonically to zero for a finite value of $k = \hat{k}$. In the worst case this will happen for $\hat{k} = n$.

Stage 3

The last part of FGMRES is the computation of $\bar{x}_k = fl(x_0 + \bar{Z}\bar{y}_k)$. The value $\bar{x}_k$ satisfies the relations

$$\begin{cases} \bar{x}_k = \bar{x}_0 + \bar{Z}_k \bar{y}_k + \delta x_k, \\ ||\delta x_k|| \leq c_3(k,1)\varepsilon \, ||\bar{Z}_k|| \, ||\bar{y}_k|| + \varepsilon \, ||\bar{x}_0|| + \mathcal{O}(\varepsilon^{\,2}). \end{cases} \tag{A.38}$$

Therefore, we have

$$\alpha_k = ||r_0 + \delta r_0 + A\delta x_k - A\delta x_k - A\bar{Z}_k \bar{y}_k - A\hat{Z}_k \bar{y}_k||. \tag{A.39}$$

From (A.38), we have

$$\begin{cases} \alpha_k = ||b - A\bar{x}_k + w|| \\ w = \delta r_0 + A\delta x_k - A\hat{Z}_k \bar{y}_k \end{cases} \tag{A.40}$$

and then

$$||b - A\bar{x}_k|| \leq ||w|| + \alpha_k. \tag{A.41}$$

From (A.34), (A.35), (A.38), (A.40), and (A.41), we have

$$\begin{cases} ||\delta r_0|| \leq c_{24}(n,1)\varepsilon \, (||b|| + ||A|| \, ||\bar{x}_0||) + \mathcal{O}(\varepsilon^{\,2}) \\ ||A\delta x_k|| \leq c_{25}(n,1)\varepsilon \, ||A|| \left[\, ||\bar{Z}_k|| \, ||\bar{y}_k|| + ||\bar{x}_0|| \right] + \mathcal{O}(\varepsilon^{\,2}) \end{cases} \tag{A.42}$$

and, finally,

$$\begin{aligned} ||w|| \quad \leq \quad & c_{26}(n,k)\varepsilon \Big(||b|| + ||A|| \, ||\bar{x}_0|| + ||\bar{H}_k|| \, ||\bar{y}_k|| + \\ & ||A|| \, ||\bar{Z}_k|| \, ||\bar{y}_k|| \Big) + \mathcal{O}(\varepsilon^{\,2}). \end{aligned} \tag{A.43}$$

Therefore, under hypothesis (5.13) and from (A.41), we have

$$\begin{aligned} ||b - A\bar{x}_k|| \quad \leq \quad & \alpha_k + c_2(n,k)\varepsilon \Big(||b|| + ||A|| \, ||\bar{x}_0|| + ||\bar{H}_k|| \, ||\bar{y}_k|| + \\ & ||A|| \, ||\bar{Z}_k|| \, ||\bar{y}_k|| \Big) + \mathcal{O}(\varepsilon^{\,2}) \end{aligned} \tag{A.44}$$

and, then, taking into account that, from (A.34),

$$||\bar{H}_k|| \leq ||A\bar{Z}_k|| + \mathcal{O}(\varepsilon \,) \; \forall k < \hat{k}, \tag{A.45}$$

we have 5.16.

From the analysis of Björck and Paige (1992) and Giraud and Langou (2002), the Gram-Schmidt orthogonalization process applied to $fl(C^{(k)})$ computes matrices $\bar{V}_k = fl(V_k)$ that lose orthogonality quickly. Nevertheless, under the mild hypothesis (5.15) on the matrix $C^{(k)}$ the spectral condition number of each $\bar{V}_k$ is close to one and

$$1 \leq ||\bar{V}_k|| \leq k$$
$$\kappa(\bar{V}_k) \leq 1.3$$

so that

$$||\bar{V}_k^+|| \leq 1.3/||\bar{V}_k|| \leq 1.3$$

where $\bar{V}_k^+ = (\bar{V}^T\bar{V}_k)^{-1}\bar{V}_k^T$, (Giraud and Langou, 2002, Thm 3.1 and Paige et al., 2006, Thm 5.2).

Taking into account the previous relations, it is possible to express $\bar{y}_k$ as a function of $\bar{x}_k - x_0$. For each column $\bar{z}_i$ of $\bar{Z}_k$ a matrix $\mathcal{G}_i$ exists and satisfies

$$(M + \mathcal{G}_i)\bar{z}_i = \bar{v}_i, \qquad ||\mathcal{G}_i|| \leq c_{27}(n,1)\,\varepsilon\,||\,|\hat{L}|\,|\hat{D}|\,|\hat{L}^T|\,|| + \mathcal{O}(\varepsilon^2),$$

where $\bar{v}_i$ is the $i-$th column of $\bar{V}_k$. Then, we have the relation

$$M\bar{Z}_k = \bar{V}_k + \hat{W}_k + \mathcal{O}(\varepsilon^2), \tag{A.46}$$

From the relations (A.38) and (A.46), it follows that

$$\begin{aligned}
&\bar{x}_k - \bar{x}_0 - \delta x_k = \bar{Z}_k\bar{y}_k \\
&\bar{V}_k^+ M(\bar{x}_k - \bar{x}_0 - \delta x_k) = \bar{V}_k^+ M\bar{Z}_k\bar{y}_k \\
&\bar{V}_k^+(M(\bar{x}_k - \bar{x}_0) - M\delta x_k) = (\bar{V}_k^+\bar{V}_k\bar{y}_k + \bar{V}_k^+\hat{W}_k\bar{y}_k) + \mathcal{O}(\varepsilon^2) \\
&\bar{V}_k^+(M(\bar{x}_k - \bar{x}_0) - M\delta x_k) - \bar{V}_k^+\hat{W}_k\bar{y}_k + \mathcal{O}(\varepsilon^2) = \bar{y}_k.
\end{aligned}$$

Then, we have

$$||\bar{y}_k|| \leq 1.3||M(\bar{x}_k - \bar{x}_0)|| + 1.3||\hat{W}_k||\,||\bar{y}_k|| + ||M||\,||\delta x_k|| + \mathcal{O}(\varepsilon^2)$$

and

$$\begin{aligned}
||\bar{y}_k|| \;&\leq\; 1.3||M(\bar{x}_k - \bar{x}_0)|| + 1.3||\hat{W}_k||\,||\bar{y}_k|| \\
&+\; c_3(k,1)\varepsilon\,||M||\,||\bar{Z}_k||\,||\bar{y}_k|| + \varepsilon\,||\bar{x}_0|| + \mathcal{O}(\varepsilon^2).
\end{aligned}$$

From hypothesis (5.17), we can state the inequality

$$||\bar{y}_k|| \leq \gamma||M(\bar{x}_k - \bar{x}_0)|| + \gamma\varepsilon\,||\bar{x}_0|| + \mathcal{O}(\varepsilon^2), \quad \text{where } \gamma = \frac{1.3}{1-\rho}. \tag{A.47}$$

Finally, by substituting (A.47) in (A.44) and taking into account (A.45), we have the following upper bound for the residual

$$\begin{aligned}
||b - A\bar{x}_k|| \;\leq\; &\alpha_k + c_4(n,k)\gamma\varepsilon\,\Big[||b|| + ||A||\,||\bar{x}_0|| + \\
&\Big(||A\bar{Z}_k|| + ||A||\,||\bar{Z}_k||\Big)\Big(||M(\bar{x}_k - \bar{x}_0)|| + \varepsilon\,||\bar{x}_0||\Big)\Big] + \mathcal{O}(\varepsilon^2).
\end{aligned}$$

Furthermore, hypothesis (5.13) implies that there exists a value $\hat{k}$ (possibly equal to n in the worst case) for which $\alpha_k = 0$. Then, we can conclude that for $k \geq \hat{k}$ we have (5.18).

B : Proof of Theorem 5.2

We can bound $||M(\bar{x}_k - \bar{x}_0)||$ using (5.19) and (5.20). Under hypothesis (5.21), we have

$$\begin{aligned} ||M(\bar{x}_k - \bar{x}_0)|| &\leq ||A\bar{x}_k - b|| + \tau\,||\bar{x}_k|| + ||\delta A||\,||\bar{x}_k|| + ||\delta M||\,||\bar{x}_0|| \\ &\leq ||A\bar{x}_k - b|| + 2\tau\big(\,||\bar{x}_k|| + ||\bar{x}_0||\,\big). \end{aligned} \tag{B.48}$$

From the relations (B.48) and (5.18), we have

$$\begin{aligned} ||b - A\bar{x}_k|| \;\leq\; & c_4(n,k)\gamma\varepsilon\,(||b|| + ||A||\,||\bar{x}_0|| + \\ & ||A||\,||\bar{Z}_k||\,\big(||A\bar{x}_k - b|| + \tau\,(\,||\bar{x}_k|| + ||\bar{x}_0||\,)\big) + \mathcal{O}(\varepsilon^{\,2}). \end{aligned}$$

Under the hypotheses (5.21) and (5.22), we have

$$\left\{ \begin{array}{rcl} ||b - A\bar{x}_k|| & \leq & 2\mu\varepsilon\,\Big(||b|| + ||A||\,||\bar{x}_0|| + \\ & & \qquad \tau||A||\,||\bar{Z}_k||\,(||\bar{x}_k|| + ||\bar{x}_0||)\Big) + \mathcal{O}(\varepsilon^{\,2}) \\ \mu = \dfrac{c_7(n,k)\gamma}{1 - c_7(n,k)\varepsilon\,\gamma||A||\,||\bar{Z}_k||}. & & \end{array} \right. \tag{B.49}$$

Under this final hypothesis (5.23), we obtain (5.24), i.e. the normwise backward stability for FGMRES using as preconditioner the M computed by `MA57` with static pivoting.

C : Proof of Theorem 5.3

The first two stages of GMRES roundoff error analysis are the same as the first two stages in the proof of Theorem 5.1. Under the assumption of performing few steps of iterative refinement in solving the final linear system

$$Mq = fl(\bar{V}_k\bar{y}_k),$$

we have the following relations:

$$\left\{ \begin{array}{l} (M + \delta M)\big(I + \Gamma\big)^{-1}\,(\bar{x}_k - (I + \Gamma)\,\bar{x}_0) = \bar{V}_k\,\bar{y}_k + \widetilde{f} \\ |\Gamma| \leq \varepsilon\,I, \\ ||\widetilde{f}|| \leq k\varepsilon\;||\bar{V}_k||\,||\bar{y}_k|| + \mathcal{O}(\varepsilon^{\,2}) \leq 1.3k\varepsilon\,||\bar{y}_k|| + \mathcal{O}(\varepsilon^{\,2}), \\ ||\delta M|| \leq n\varepsilon\;||M|| + \mathcal{O}(\varepsilon^{\,2}). \end{array} \right. \tag{C.50}$$

Taking into account that

$$(M + \delta M)\,(I + \Gamma)^{-1} = M + \Big(\delta M - M\,\Gamma\Big)\Big(I + \Gamma\Big)^{-1} = M + \widetilde{\delta M}$$

with

$$||\widetilde{\delta M}|| \leq (n+1)\frac{\varepsilon}{1-\varepsilon}||M|| = c_{28}(n,1)\varepsilon\,||M||,$$

we have, from (C.50), that

$$(M + \widetilde{\delta M})(\bar{x}_k - \bar{x}_0) = \bar{V}_k\,\bar{y}_k + \widetilde{f} + (M + \widetilde{\delta M})\,\Gamma\,\bar{x}_0. \tag{C.51}$$

Therefore, we have from (5.25) and (C.51) that (A.38) can be replaced by

$$\begin{aligned} \bar{x}_k &= \bar{x}_0 + (M + \widetilde{\delta M})^{-1}(\bar{V}_k\,\bar{y}_k + \widetilde{f}) + \Gamma\,\bar{x}_0 && \text{(C.52)}\\ &= \bar{x}_0 + M^{-1}\,\bar{V}_k\,\bar{y}_k + \delta\bar{x}_k && \text{(C.53)} \end{aligned}$$

with

$$\begin{aligned} \delta\bar{x}_k &= \Gamma\,\bar{x}_0 + M^{-1}\,\widetilde{f} - M^{-1}\,\widetilde{\delta M}\,M^{-1}\,\bar{V}_k\,\bar{y}_k + \mathcal{O}(\varepsilon^{\,2})\\ &= \Gamma\,\bar{x}_0 + M^{-1}\,\widetilde{f} - M^{-1}\,\widetilde{\delta M}\,(\bar{x}_k - \bar{x}_0 - \delta\bar{x}_k) + \mathcal{O}(\varepsilon^{\,2}) \end{aligned}$$

and

$$\begin{aligned} \delta\bar{x}_k &= \big(I - M^{-1}\,\widetilde{\delta M}\big)^{-1}\big(\Gamma\,\bar{x}_0 + M^{-1}\,\widetilde{f} - M^{-1}\,\widetilde{\delta M}\,(\bar{x}_k - \bar{x}_0)\big) + \mathcal{O}(\varepsilon^{\,2})\\ &= \Gamma\,\bar{x}_0 + M^{-1}\,\widetilde{f} - M^{-1}\,\widetilde{\delta M}\,(\bar{x}_k - \bar{x}_0) + \mathcal{O}(\varepsilon^{\,2}). \end{aligned} \qquad \text{(C.54)}$$

Under hypotheses (5.15) and (5.25), we obtain from (C.52), (C.53), (C.54), and (A.46)

$$\left\{\begin{array}{lll} \alpha_k &= ||b - A\bar{x}_k + w_g|| \qquad \text{where}\\ w_g &= \delta r_0 + A\delta x_k - A\hat{Z}_k\bar{y}_k + A(M^{-1}\,\bar{V}_k\,\bar{y}_k - \bar{Z}_k\,\bar{y}_k)\\ &= \delta r_0 + A\delta x_k - A\hat{Z}_k\bar{y}_k + AM^{-1}\,\hat{W}_k\,\bar{y}_k + \mathcal{O}(\varepsilon^{\,2}), \end{array}\right. \qquad \text{(C.55)}$$

and, then

$$||b - A\bar{x}_k|| \le \alpha_k + ||w_g||. \qquad \text{(C.56)}$$

From (A.34), (A.35), (A.37), (C.52), (C.54), (C.55), and (C.56), we have

$$\left\{\begin{array}{rcl} ||\delta r_0|| &\le& c_{29}(n,1)\varepsilon\,(||b|| + ||A||\,||\bar{x}_0||) + \mathcal{O}(\varepsilon^{\,2})\\ ||A\delta x_k|| &\le& c_{30}(n,1)\varepsilon\,\Big[||AM^{-1}||\,\Big(||M||\,||\bar{x}_k - \bar{x}_0|| + ||\bar{y}_k||\Big) +\\ & & ||A||\,||\bar{x}_0||\Big] + \mathcal{O}(\varepsilon^{\,2}) \end{array}\right. \qquad \text{(C.57)}$$

and, finally,

$$\begin{aligned} ||w_g|| \le\ & c_{31}(n,k)\varepsilon\,\Big[\,||b|| + ||A||\,||\bar{x}_0|| + ||\bar{H}_k||\,||\bar{y}_k|| + ||A||\,||\bar{Z}_k||\,||\bar{y}_k|| +\\ & ||AM^{-1}||\,\Big(||M||\,||\bar{x}_k - \bar{x}_0|| + ||\bar{y}_k|| +\\ & ||\,|\hat{L}|\,|\hat{D}|\,|\hat{L}^T|\,||\,||\bar{y}_k||\Big)\,\Big] + \mathcal{O}(\varepsilon^{\,2}). \end{aligned} \qquad \text{(C.58)}$$

From (C.51), we can compute an upper bound for $||\bar{y}_k||$ by multiplying the equation by $\bar{V}_k^+$:

$$\begin{aligned} \bar{y}_k &= \bar{V}_k^+\,(M + \widetilde{\delta M})(\bar{x}_k - \bar{x}_0) - \bar{V}_k^+\,(\widetilde{f} + (M + \widetilde{\delta M})\,\Gamma\,\bar{x}_0) && \text{(C.59)}\\ &= \bar{V}_k^+\,M(\bar{x}_k - \bar{x}_0) - \bar{V}_k^+\,(\widetilde{f} - \widetilde{\delta M}(\bar{x}_k - \bar{x}_0) + (M + \widetilde{\delta M})\,\Gamma\,\bar{x}_0), && \text{(C.60)} \end{aligned}$$

and , then,

$$\begin{aligned} ||\bar{y}_k|| &\le 1.3||M(\bar{x}_k - \bar{x}_0)|| + 1.3k\varepsilon\,||\bar{y}_k|| +\\ & \quad 1.3n\varepsilon\,||M||\,(||\bar{x}_k - \bar{x}_0|| + ||\bar{x}_0||) + \mathcal{O}(\varepsilon^{\,2}). \end{aligned} \qquad \text{(C.61)}$$

From hypothesis (5.15), we can assume that $1.3\,k\varepsilon < 1$, and thus we have

$$\begin{cases} ||\bar{y}_k|| \leq \chi ||M(\bar{x}_k - \bar{x}_0)|| + \chi n\varepsilon\, ||M||\, (||\bar{x}_k - \bar{x}_0|| + ||\bar{x}_0||) + \mathcal{O}(\varepsilon^{\,2}) \\ \chi = \dfrac{1.3}{1 - 1.3k\varepsilon}. \end{cases} \tag{C.62}$$

Furthermore, as in the proof of Theorem 5.1, hypothesis (5.13) implies that there exists a value $\hat{k}$ (possibly equal to n in the worst case) for which $\alpha_k = 0$. Then, under hypotheses (5.13), (5.15), and from (C.62), (C.56), and (C.58), we can conclude that for $k \geq \hat{k}$ we have (5.26).