

# Convergence in backward error of relaxed GMRES

L. Giraud\*    S. Gratton\*    J. Langou†

CERFACS Technical Report TR/PA/04/132  
December 2004

## Abstract

This work is the follow-up of the experimental study presented in [3]. It is based on and extends some theoretical results in [15, 18]. In a backward error framework we study the convergence of GMRES when the matrix-vector products are performed inaccurately. This inaccuracy is modeled by a perturbation of the original matrix. We prove the convergence of GMRES when the perturbation size is proportional to the inverse of the computed residual norm; this implies that the accuracy can be relaxed as the method proceeds which gives rise to the terminology relaxed GMRES. As for the exact GMRES we show under proper assumptions that only happy breakdowns can occur. Furthermore the convergence can be detected using a by-product of the algorithm. We explore the links between relaxed right-preconditioned GMRES and flexible GMRES. In particular this enables us to derive a proof of convergence of FGMRES. Finally we report results on numerical experiments to illustrate the behaviour of the relaxed GMRES monitored by the proposed relaxation strategies.

## 1 Introduction

We consider the solution of a linear system of equations  $Ax = b$  using the GMRES [14] iterative method, where  $A$  is a nonsingular  $n \times n$  matrix. In some applications, performing inexact matrix-vector products in this method may be interesting as long as the convergence of GMRES is maintained. Many situations in scientific computing may benefit from such a scheme. For instance a natural application of this idea occurs in computational electromagnetic, where the fast multipole method provides approximations of the matrix-vector product within a user-defined accuracy; the less accurate the matrix-vector, the faster the computation. The key point is then to design a criterion to control the accuracy of the matrix-vector product so that the iterates achieve a satisfactory convergence

---

\*CERFACS, 42 av. Gaspard Coriolis, 31057 Toulouse cedex 1, France.

†Department of Computer Science, University of Tennessee, Knoxville, Tennessee, USA

level. Another example arises in non-overlapping domain decomposition where the matrix-vector involving the Schur complement can be approximated.

In [3], a criterion is proposed for general systems and its numerical behaviour is illustrated on a large set of numerical experiments. This work is based on some heuristic considerations and the approach is referred to as relaxation strategy because the perturbation size can grow as the inverse of the residual norm. We denote by relaxed GMRES an inexact GMRES that implements a relaxation strategy. The relaxation strategy proposed in [3] attempts to ensure the convergence of the GMRES iterates  $x_k$  within a relative normwise backward error

$$\begin{aligned} \eta_{A,b}(x_k) &= \min_{\Delta A, \Delta b} \{ \tau > 0 : \|\Delta A\| \leq \tau \|A\|, \|\Delta b\| \leq \tau \|b\| \\ &\quad \text{and } (A + \Delta A)x_k = b + \Delta b \} \\ &= \frac{\|Ax_k - b\|}{\|A\| \|x_k\| + \|b\|} \end{aligned} \quad (1)$$

less than a prescribed quantity  $\varepsilon > 0$ . A significant number of numerical experiments are presented that illustrate the merits of the strategy but also reveal its lack of robustness; for a few examples the convergence at  $\varepsilon$  is not obtained. Nevertheless, based on [3] similar strategies have been successfully applied to the solution of heterogeneous diffusion problems using domain decomposition [4], to the preconditioning of a radiation diffusion problem [19], to the solution of electromagnetic problems [11], in lattice quantum chromodynamics [5] and in ocean circulation model for steady barotropic flow [10]. A significant step toward a theoretical explanation of this observed behaviour is proposed in [15, 18]. In these latter works, important justifications are brought to the fact that under some assumptions relaxed GMRES converges in residual norm.

The convergence of iterative solvers is often based on normwise backward error criteria [2, 6, 8]. In this paper we propose criteria to control the accuracy of the matrix-vector products and show that they ensure the convergence of GMRES with respect either to  $\eta_{A,b}(x_k)$  defined in (1) or to

$$\begin{aligned} \eta_b(x_k) &= \min_{\Delta b} \{ \tau > 0 : \|\Delta b\| \leq \tau \|b\| \text{ and } Ax_k = b + \Delta b \} \\ &= \frac{\|Ax_k - b\|}{\|b\|}. \end{aligned} \quad (2)$$

We mention that  $\eta_{A,b}$  and  $\eta_b$  are recommended in [2] when the concern related to the stopping criterion is discussed.

The forthcoming development shows that deriving a sufficient convergence condition of relaxed GMRES for the stopping criterion  $\eta_b(x_k) \leq \varepsilon$  is far simpler than for  $\eta_{A,b}(x_k) \leq \varepsilon$ . However we believe that both stopping criteria are of practical interest. Both may be used to account for data uncertainties. It is problem and application dependent to decide whether these uncertainties come chiefly from the right-hand side, then  $\eta_b$  must be used, or from the matrix and the right-hand side, then  $\eta_{A,b}$  must be used.

The paper is organized as follows. In Section 2 we study the convergence of relaxed GMRES. We first recall and discuss in Section 2.1 some theoretical results established in [15, 18]. Section 2.2 is devoted to the convergence proofs. In particular we show that under suitable assumptions only happy breakdowns can occur and we explain how the perturbation size can be monitored to ensure the convergence of relaxed GMRES with respect to either  $\eta_{A,b}$  or  $\eta_b$ . We provide a stopping criterion that only uses by-products of the algorithm and does not require an exact product by  $A$  to compute the backward errors. Numerical experiments that illustrate these theoretical results are given in Section 2.3. In Section 3, we study the situation when the preconditioner is perturbed. This might happen for instance when the application of the preconditioner is obtained by solving iteratively an auxiliary linear system. We also exploit the links between relaxed preconditioned GMRES and flexible GMRES [12] and we derive a straightforward proof of convergence of FGMRES. In Section 4 we consider the use of these strategies in a restarted framework. Finally we conclude with some comments in Section 5.

In this paper, the 2-norm of a vector  $x$  is denoted by  $\|x\|$ , the spectral 2-norm of a matrix  $A$  is denoted by  $\|A\|$ . We use the notation  $\sigma_{\max}(A)$  ( $\sigma_{\min}(A)$ ) for the smallest (resp. the largest) singular value of  $A$ . The spectral condition number of a matrix is  $\kappa(A) = \|A\|\|A^{-1}\| = \sigma_{\max}(A)/\sigma_{\min}(A)$ .

## 2 GMRES with inexact matrix-vector

### 2.1 Background and existing results

The GMRES method is based on the Arnoldi recursion  $AV_k = V_{k+1}\bar{H}_k$ , where  $V_k = [v_1, \dots, v_k]$  is an orthogonal matrix, and  $\bar{H}_k$  is a  $(k+1) \times k$  upper-Hessenberg matrix. We assume that we are given an initial guess  $x_0$  for the solution  $x^*$  of the system. If  $x_0 \neq x^*$ , the GMRES algorithm generates a sequence of iterates  $\{x_k\}_{k=1,2,\dots}$  such that  $x_k$  realizes the minimum of the 2-norm of the residual  $r_k = b - Ax_k$  over the space  $x_0 + \mathcal{K}_k(A, v_1)$ . The Krylov subspace  $\mathcal{K}_k(A, v_1)$  is defined by  $\mathcal{K}_k(A, v_1) = \text{span}(v_1, Av_1, \dots, A^{k-1}v_1)$  where  $\beta v_1 = r_0$  and  $\beta = \|r_0\| \neq 0$ . From the unitary invariance of the 2-norm,  $x_k = x_0 + V_k y_k$  where  $y_k$  is the solution of the linear least-squares problem  $\min_y \|\bar{H}_k y - \beta e_1\|$ , where  $e_1$  is the first vector of the canonical basis.

We assume that it is possible to monitor the accuracy of the matrix-vector product  $Av$  of the Arnoldi procedure. From a mathematical point of view, the inaccuracy can be modeled by introducing a perturbation matrix  $E$ , depending possibly on  $v$ , such that  $(A+E)v$  is the quantity actually computed. At step  $k$  of this perturbed Arnoldi algorithm, the vector  $w = (A + E_k)v_k$  is orthogonalized against the vectors  $v_j$ ,  $j = 1, \dots, k$  so that the following relation holds,

$$[(A + E_1)v_1, \dots, (A + E_k)v_k] = [v_1, \dots, v_k, v_{k+1}]\bar{H}_k. \quad (3)$$

Note that this matrix perturbation approach was taken in [3, 15]. In [18] the inaccuracies are modeled by introducing the vector  $f$  such that  $w = Av + f$ .

We took the former approach because it generalizes to the latter ones by setting  $E = fv^T$  since  $v^T v = 1$ . In the above equality (3), referred to as inexact Arnoldi in [15],  $V_k = [v_1, \dots, v_k]$  is an orthogonal matrix and  $\bar{H}_k$  is a  $(k+1) \times k$  upper-Hessenberg matrix. Similarly to exact GMRES, the  $k$ -th iterate of the inexact method is defined by  $x_k = x_0 + V_k y_k$  where  $y_k$  is the solution of the linear least-squares problem  $\min_y \|\bar{H}_k y - \beta e_1\|$ . Introducing the perturbation matrix  $G_k = [E_1 v_1, \dots, E_k v_k]$  the inexact Arnoldi relation can also be written [18] as an exact Arnoldi relation  $\tilde{A}_k V_k = (A + G_k V_k^T) V_k = V_{k+1} \bar{H}_k$ , with  $\tilde{A}_k = A + G_k V_k^T$ . This last equality shows that the quantities  $x_i, \bar{H}_i$  and  $v_i$  for  $i \leq k$  generated by the inexact GMRES until step  $k$  are the same as those generated by exact GMRES applied to the linear system  $\tilde{A}_k x = b$  until step  $k$ . Using classical results on GMRES [13] this observation implies by induction that the norm of the *computed residual*  $\tilde{r}_k = b - \tilde{A}_k x_k$  is monotonically decreasing as  $k$  grows. The inexact Arnoldi recursion can also be written

$$\tilde{A}_k V_k = V_k H_k + h_{k+1,k} v_{k+1} e_k.$$

Using an analogy with the terminology used for the exact GMRES algorithm, we say that a *breakdown* occurs at step  $m$  if  $h_{m+1,m} = 0$ . In what follows, the step  $m$  refers to the step where the breakdown occurs. Because of the orthogonality of  $V_k$ , similarly to the exact GMRES framework, such a breakdown certainly occurs for  $m \leq n$ . At each step the residual gap is defined by  $r_k - \tilde{r}_k$  where  $r_k = b - Ax_k$ . Let  $\epsilon$  be a positive real number. Assuming that the inexact method can be run until step  $\ell$  without breakdown, it is shown in [15, Theorem 5.3] that if

$$\|E_k\| \leq \frac{\sigma_{\min}(\bar{H}_\ell)}{\ell} \frac{\epsilon}{\|\tilde{r}_{k-1}\|} \quad (4)$$

for  $k \leq \ell$  then  $\|r_\ell - \tilde{r}_\ell\| \leq \epsilon$ . Because the norm of the computed residual,  $\|\tilde{r}_k\|$ , is monotonically decreasing as  $k$  grows, the inequalities (4) and

$$\|r_k\| \leq \|r_k - \tilde{r}_k\| + \|\tilde{r}_k\| \quad (5)$$

ensure that the residual norm satisfies at step  $\ell$

$$\|r_\ell\| \leq \epsilon + \|\tilde{r}_\ell\|. \quad (6)$$

However, it can be seen that there is an implicit relation in (4) linking all the  $E_k$ 's,  $k \leq \ell$  together. The  $E_k$ 's depend on  $\bar{H}_\ell$  that itself depends on the  $E_k$ 's. This means that it is not possible to implement (4) in an algorithm.

Another interesting feature would also be to monitor  $\|\tilde{r}_\ell\|$  in (6) so that it can be less than any prescribed value for large enough  $\ell$ . In this way, the residual norm  $\|r_\ell\|$ , which is a key ingredient of the backward errors  $\eta_{A,b}(x_\ell)$  and  $\eta_b(x_\ell)$ , would also be controlled through (6). It turns out that contrary to the exact GMRES,  $\|\tilde{r}_\ell\|$  is not necessarily zero at the breakdown of the inexact algorithm. For example, for  $A = \text{diag}(1, 2)$ ,  $b = (1/\sqrt{2}, 1/\sqrt{2})^T$  and  $x_0 = (0, 0)^T$ , if we take  $E_1 = 0$  and  $E_2 = \text{diag}(-2, 0)$ , the inexact Arnoldi relation reads

$$[(A + E_1)v_1, (A + E_2)v_2] = [v_1, v_2] \begin{pmatrix} 3/2 & 3/2 \\ 1/2 & 1/2 \end{pmatrix}.$$

We therefore have a breakdown of the inexact GMRES method, but  $\|\tilde{r}_1\| = \|\tilde{r}_2\| = 1/\sqrt{10}$ . Note that  $H_2$  is singular consequently the linear least-squares problem

$\min_y \|\bar{H}_2 y - \beta e_1\|$  is rank deficient. This means that there is an infinite number of solutions to the linear least-squares problem and neither  $y_2$  nor  $x_2$  is well defined.

The purpose of the next Section is to fix the problems related to the control of  $\|\tilde{r}_\ell\|$  and to the possible singularity of  $H_m$  as well as to remove the implicit relation linking all the  $E_k$ 's,  $k \leq \ell$  together. We shall see that this will enable us to design a relaxed GMRES algorithm that is guaranteed to converge for the backward error criteria  $\eta_b \leq \varepsilon$  or  $\eta_{A,b} \leq \varepsilon$ , for any prescribed tolerance  $\varepsilon$ .

## 2.2 Theoretical results

We first design a strategy that ensures the convergence of the inexact GMRES iterates so that  $\eta_b$  can be made smaller than any prescribed tolerance; this theoretical result is closely related to the work of [15]. We consider the splitting

$$\eta_b(x_k) \leq \frac{\|r_k - \tilde{r}_k\|}{\|b\|} + \frac{\|\tilde{r}_k\|}{\|b\|}$$

and set accordingly

$$\varepsilon = \varepsilon_g + \varepsilon_c \tag{7}$$

where  $\varepsilon_g$  (resp.  $\varepsilon_c$ ) is the targeted tolerance for the scaled residual gap  $\frac{\|r_k - \tilde{r}_k\|}{\|b\|}$  (resp. the relative *computed* residual norm  $\frac{\|\tilde{r}_k\|}{\|b\|}$ ).

In the framework of the inexact GMRES the situation where the computed residual  $\tilde{r}_m$  is zero at the breakdown is referred to as a happy breakdown. Such a situation is of interest because it ensures that  $\frac{\|\tilde{r}_k\|}{\|b\|}$  can be made smaller than any prescribed  $\varepsilon_c$  provided that  $k$  is large enough. The following Theorem shows a possible way to control the  $E_k$ 's such that only a happy breakdown can eventually occur.

### Theorem 1 [Breakdown in inexact GMRES]

*The following results hold for the inexact GMRES.*

1. *Suppose that  $r_0 \neq 0$ , that  $h_{k+1,k} \neq 0$  for  $k < m$ . If  $H_m$  is nonsingular, the computed residual  $\tilde{r}_m$  is zero iff  $h_{m+1,m}$  is zero.*

2. *Let  $c$  be such that  $0 < c < 1$ . If for any  $k$ ,  $\|E_k\| \leq c \frac{\sigma_{\min}(A)}{n}$  then  $(1-c)\sigma_{\min}(A) \leq \sigma_{\min}(\bar{H}_k)$ . This latter inequality implies that  $H_m$  is nonsingular and that only a happy breakdown can occur.*

**Proof:** 1. The inexact Arnoldi relation (3) can also be written:

$$\begin{aligned} [(A + E_1)v_1, \dots, (A + E_k)v_k] &= [v_1, \dots, v_k, v_{k+1}] \bar{H}_k \\ A[(I + A^{-1}E_1)v_1, \dots, (I + A^{-1}E_k)v_k] &= [v_1, \dots, v_k, v_{k+1}] \bar{H}_k. \end{aligned}$$

The last equation reads as Flexible GMRES relation, where the variable preconditioner at step  $j \leq k$  is  $M_j = (I + A^{-1}E_j)$ . We are then in situation to apply [13, Proposition 9.3] which concludes the proof.

2. The orthogonality of  $V_k$  and the relaxed Arnoldi relation reads  $V_{k+1}\bar{H}_k = (A + G_k V_k^T)V_k$ , from which follows that

$$\sigma_{\min}(A + G_k V_k^T) \leq \sigma_{\min}(\bar{H}_k).$$

Assuming that there exists a constant  $c$ ,  $0 < c < 1$  such that

$$\|E_k\| \leq c \frac{\sigma_{\min}(A)}{n} \quad (8)$$

we get

$$\|G_k V_k^T\| = \left\| \sum_{i=1}^m E_i v_i v_i^T \right\| \leq \sum_{i=1}^m \|E_i v_i v_i^T\| \leq c \sigma_{\min}(A), \quad (9)$$

and inequality  $\sigma_{\min}(A) - \|G_k V_k^T\| \leq \sigma_{\min}(A + G_k V_k^T)$  (see [1, Theorem 3.3.16]) implies that  $0 < (1 - c)\sigma_{\min}(A) \leq \sigma_{\min}(\bar{H}_k)$ . This inequality indicates that  $\bar{H}_k$  has full rank. In particular this is true at breakdown which shows that  $H_m$  is nonsingular. □

The next theorem gives a sufficient condition for the convergence of relaxed GMRES for  $\eta_b$ .

**Theorem 2** [Convergence of relaxed GMRES for  $\eta_b$ ]

Let  $c$  be such that  $0 < c < 1$  and let  $\varepsilon_c$  and  $\varepsilon_g$  be any positive real numbers. Assume for all  $k$  that

$$\|E_k\| \leq \frac{c}{n} \sigma_{\min}(A) \min \left( 1, \frac{\|b\|}{\|\tilde{r}_{k-1}\|} \varepsilon_g \right). \quad (10)$$

There exists  $\ell$ ,  $0 < \ell \leq n$ , such that the following stopping criterion is satisfied

$$\frac{\|\tilde{r}_\ell\|}{\|b\|} \leq \varepsilon_c \quad (11)$$

and

$$\eta_b(x_\ell) \leq \varepsilon_c + \varepsilon_g.$$

Proof: Substituting  $\|b\|\varepsilon_g$  for  $\epsilon$  in (4) yields  $\frac{\|r_\ell - \tilde{r}_\ell\|}{\|b\|} \leq \varepsilon_g$ . The conclusion follows from (5),  $\|\tilde{r}_\ell\| \leq \varepsilon_c \|b\|$  and the second statement of Theorem 1. □

Some comments on the above result are in order. The stopping criterion (11) is based on  $\|\tilde{r}_k\| = |h_{\ell+1,\ell}|$  which is a by-product of the algorithm and does not require any additional matrix-vector product. Similarly the control on  $\|E_k\|$  only

involves some constants of the problem ( $n$ ,  $\sigma_{\min}(A)$ ,  $\|b\|$ ) and the by-product  $\|\tilde{r}_k\|$ . Finally, the constant  $c$  can be any value between zero and one.

We now show a similar result for a prescribed accuracy on the backward error  $\eta_{A,b}$ . In a first step, using the same technique as for Theorem 2, we obtain in the following lemma a control that ensures the convergence with respect to  $\eta_{A,b}$ .

**Lemma 1** *Let  $\varepsilon_c$  and  $\varepsilon_g$  be any positive real numbers. Suppose that for all  $k$ ,*

$$\|E_k\| \leq \frac{\sigma_{\min}(\bar{H}_\ell) \|A\| \|x_\ell\| + \|b\|}{n \|\tilde{r}_{k-1}\|} \varepsilon_g \quad (12)$$

and that  $\|\tilde{r}_\ell\| \leq \varepsilon_c \|A\| \|x_\ell\|$  then

$$\eta_{A,b}(x_\ell) \leq \varepsilon_c + \varepsilon_g.$$

**Proof:** Substituting  $(\|A\| \|x_k\| + \|b\|) \varepsilon_g$  for  $\epsilon$  in (4) yields  $\frac{\|r_\ell - \tilde{r}_\ell\|}{\|A\| \|x_k\| + \|b\|} \leq \varepsilon_g$ . The conclusion follows from (5) and  $\|\tilde{r}_\ell\| \leq \varepsilon_c (\|A\| \|x_\ell\| + \|b\|)$ . □

The above result is not implementable because of the forward reference to the quantities  $\|x_\ell\|$  and  $\sigma_{\min}(\bar{H}_\ell)$ . The next lemma shows that the control ensuring a happy breakdown  $\|r_m\| = 0$ , i.e.  $\|E_k\| \leq c \frac{\sigma_{\min}(A)}{n}$ , also enables us to derive a lower-bound on  $\|x_\ell\|$  when  $\|\tilde{r}_\ell\| \leq \varepsilon_c \|A\| \|x_\ell\|$  is satisfied. This lower bound is expressed in terms of  $c$ ,  $\varepsilon_c$  and some constants of the problem.

**Lemma 2** *Let  $c$  and  $\varepsilon_c$  be such that  $0 < c < \frac{1}{2}$  and  $0 < \varepsilon_c < \frac{1 - 2c}{2\kappa(A)} \frac{\kappa(A) - c}{\kappa(A) + c}$ .*

*Assume  $\|E_k\| \leq c \frac{\sigma_{\min}(A)}{n}$ , for all  $k \leq \ell$  and  $\frac{\|\tilde{r}_\ell\|}{\|A\| \|x_\ell\|} \leq \varepsilon_c$ , then*

$$0 \leq \frac{1 - 2\omega}{1 - \omega} \|x^*\| \leq \|x_\ell\| \quad (13)$$

where  $\omega = c + \varepsilon_c \kappa(A) \frac{\kappa(A) + c}{\kappa(A) - c}$  is such that  $0 < \omega < 1/2$ .

**Proof:** From the assumptions of the theorem, we have

$$\frac{\|\tilde{r}_\ell\|}{\|\tilde{A}\| \|x_\ell\|} \leq \varepsilon_c \frac{\|A\|}{\|\tilde{A}\|}$$

and using (9),

$$\|\tilde{A}\| = \|A + GV_\ell^T\| \geq \|A\| - \|GV_\ell^T\| \geq \|A\| - c\sigma_{\min}(A). \quad (14)$$

Similarly we have the upper bound

$$\|\tilde{A}\| \leq \|A\| + c\sigma_{\min}(A). \quad (15)$$

The inequality (14) implies that

$$\frac{\|\tilde{r}_\ell\|}{\|\tilde{A}\|\|x_\ell\|} \leq \varepsilon_c \frac{\|A\|}{\|A\| - c\sigma_{\min}(A)}, \quad (16)$$

because  $0 < c < 1/2$ . In a backward error framework the iterate  $x_\ell$  can be viewed as the solution of the perturbed linear system  $(\tilde{A}_\ell + \Delta\tilde{A})x_\ell = b$  with, from (16) and (1) give

$$\|\Delta\tilde{A}\| \leq \varepsilon_c \frac{\|A\|}{\|A\| - c\sigma_{\min}(A)} \|\tilde{A}\|. \quad (17)$$

Combining (15) and (17) leads to

$$\|\Delta\tilde{A}\| \leq \varepsilon_c \|A\| \frac{\kappa(A) + c}{\kappa(A) - c}. \quad (18)$$

The iterate  $x_\ell$  can also be viewed as the solution of  $(A + GV_\ell^T + \Delta\tilde{A})x_\ell = b = (A + E)x_\ell$ , with

$$\|E\| = \|GV_\ell^T + \Delta\tilde{A}\| \leq \|GV_\ell^T\| + \|\Delta\tilde{A}\| \leq c\sigma_{\min}(A) + \varepsilon_c \|A\| \frac{\kappa(A) + c}{\kappa(A) - c}. \quad (19)$$

We can now apply [9, Theorem 7.2] to the linear system  $Ax = b$  with perturbation  $\|E\|$  in the following form: if

$$\|A^{-1}\|\|E\| < \frac{1}{2} \quad (20)$$

then

$$\|x^* - x_\ell\| = \frac{\|A^{-1}\|\|E\|}{1 - \|A^{-1}\|\|E\|} \|x^*\|. \quad (21)$$

Using (19), (20) holds if  $\omega < 1/2$ , which is the case if

$$\varepsilon_c < \frac{1 - 2c}{2\kappa(A)} \frac{\kappa(A) - c}{\kappa(A) + c}. \quad (22)$$

Finally, combining  $\|x^*\| - \|x_\ell\| \leq \|x^* - x_\ell\|$  with (21) and (22) concludes the proof. □

The next theorem gives a sufficient convergence condition for relaxed GMRES for  $\eta_{A,b}$ . Similarly as for Theorem 10 the control on  $\|E_k\|$  and the stopping criterion involved in relaxed GMRES are based on by-products of the algorithm and some constants of the problem. Provided that the above constants are available or easily estimated, this result can be used in a practical implementation.

**Theorem 3** [Convergence of relaxed GMRES for  $\eta_{A,b}$ ]

Let  $\varepsilon_c$  and  $\varepsilon_g$  be any positive real numbers. Let  $c$  and  $\omega$  be such that  $0 < c < \frac{1}{2}$ ,  $\varepsilon_c < \frac{1-2c}{2\kappa(A)} \frac{\kappa(A)-c}{\kappa(A)+c}$ . Suppose that for all  $k$

$$\|E_k\| \leq \frac{1}{n} \sigma_{\min}(A) \min \left( c, (1-c) \frac{\gamma^*}{\|\tilde{r}_{k-1}\|} \varepsilon_g \right), \quad (23)$$

where  $\gamma^* = \frac{1-2\omega}{1-\omega} \|A\| \|x^*\| + \|b\|$  and  $\omega = c + \varepsilon_c \kappa(A) \frac{\kappa(A)+c}{\kappa(A)-c}$ . There exists  $\ell$ ,  $0 < \ell \leq n$ , such that the following stopping criterion is satisfied

$$\frac{\|\tilde{r}_\ell\|}{\|A\| \|x_\ell\|} \leq \varepsilon_c \quad (24)$$

and

$$\eta_{A,b}(x_\ell) \leq \varepsilon_c + \varepsilon_g.$$

**Proof:** We first observe that  $\|E_k\| \leq \frac{c}{n} \sigma_{\min}(A)$  implies (see Theorem 1) that only a happy breakdown may occur. Because  $\tilde{r}_k$  is monotonically decreasing there exists a  $\ell$  such that  $\frac{\|\tilde{r}_\ell\|}{\|A\| \|x_\ell\|} \leq \varepsilon_c$ . Furthermore, from the assumptions

$$\|E_k\| \leq \left( \frac{(1-c)\sigma_{\min}(A)}{n} \right) \left( \frac{1-2\omega}{1-\omega} \|A\| \|x^*\| + \|b\| \right) \frac{\varepsilon_g}{\|\tilde{r}_k\|}. \quad (25)$$

The first factor in the right-hand side of (25) can be bounded above by  $\frac{\sigma_{\min}(\bar{H}_\ell)}{n}$ , because  $\|E_k\| \leq \frac{c}{n} \sigma_{\min}(A)$  and the second statement of Theorem 1 applies. Using the conclusions of Lemma 2, the second factor is bounded above by  $\|A\| \|x_\ell\| + \|b\|$  because of (13) and because  $0 < \omega < 1/2$  implies  $0 < \frac{1-2\omega}{1-\omega}$ . Therefore we have  $\|E_k\| \leq \frac{\sigma_{\min}(\bar{H}_\ell)}{n} \frac{\|A\| \|x_\ell\| + \|b\|}{\|\tilde{r}_{k-1}\|} \varepsilon_g$ . Consequently the conditions of Lemma 1 hold which concludes the proof.  $\square$

**Remark 1** As illustrated in the forthcoming numerical experiments, it might be noticed that at the drawback of smaller perturbations one can derive a control that does not require the knowledge of  $\|x^*\|$  (or of a lower bound). If we replace  $\gamma^*$  by  $\gamma^b = \|b\| < \gamma^*$  in Theorem 3, the theorem is still valid.

So far, similarly to [15, 18], we have considered that the initial residual involved in the inexact algorithm is computed exactly :  $r_0 = b - Ax_0$ . We now provide an extension to these results which additionally accounts for inaccuracies in the very first matrix-vector product  $Ax_0$ . Such a result is crucial when only approximations of the matrix-vector product are available. This is the case for instance when  $A$  is a Schur complement matrix in domain decomposition where the local systems are solved iteratively [4].

**Theorem 4** [Convergence of relaxed GMRES with inexact initial residual]

Let  $\varepsilon_A$  and  $\varepsilon_b$  be any positive real numbers. Suppose that the initial residual  $r_0$  is approximated with  $\tilde{r}_0 = b + \Delta b - (A + \Delta A)x_0$  with  $\|\Delta A\| \leq \varepsilon_A \|A\|$  and  $\|\Delta b\| \leq \varepsilon_b \|b\|$ . Suppose that the assumptions of Theorem 3 hold. There exists  $\ell$ ,  $0 < \ell \leq n$ , such that the stopping criterion (24) is satisfied and

$$\eta_{A,b}(x_\ell) \leq \varepsilon_c + \varepsilon_g + (1 + \varepsilon_c + \varepsilon_g) \left( \varepsilon_A \frac{\|x_0\|}{\|x_\ell\|} + \varepsilon_b \right).$$

Proof: Starting inexact GMRES with the approximated residual  $\tilde{r}_0$  is equivalent to start inexact GMRES without any initial perturbations on  $A$  or  $b$  for the solution of  $Ax = b + \Delta b - \Delta Ax_0 = \tilde{b}$  with initial guess  $x_0$ .

Using Theorem 4 on that latter system leads to  $\frac{\|Ax_\ell - \tilde{b}\|}{\|A\|\|x_\ell\| + \|\tilde{b}\|} \leq \varepsilon_c + \varepsilon_g$ .

Furthermore we have

$$\begin{aligned} \frac{\|Ax_\ell - b\|}{\|A\|\|x_\ell\| + \|b\|} &\leq \frac{\|Ax_\ell - \tilde{b}\|}{\|A\|\|x_\ell\| + \|\tilde{b}\|} + \frac{\|\Delta b - \Delta Ax_0\|}{\|A\|\|x_\ell\| + \|b\|} \\ &\leq \frac{\|Ax_\ell - \tilde{b}\|}{\|A\|\|x_\ell\| + \|\tilde{b}\|} \frac{\|A\|\|x_\ell\| + \|\tilde{b}\|}{\|A\|\|x_\ell\| + \|b\|} + \frac{\|\Delta b - \Delta Ax_0\|}{\|A\|\|x_\ell\| + \|b\|} \\ &\leq (\varepsilon_c + \varepsilon_g) \left( 1 + \frac{\|\Delta b - \Delta Ax_0\|}{\|A\|\|x_\ell\| + \|\tilde{b}\|} \right) + \frac{\|\Delta b - \Delta Ax_0\|}{\|A\|\|x_\ell\| + \|b\|} \\ &\leq (\varepsilon_c + \varepsilon_g) + (1 + \varepsilon_c + \varepsilon_g) \frac{\|\Delta b - \Delta Ax_0\|}{\|A\|\|x_\ell\| + \|b\|} \end{aligned}$$

the fact that

$$\frac{\|\Delta b - \Delta Ax_0\|}{\|A\|\|x_\ell\| + \|b\|} \leq \frac{\|\Delta b\| + \|\Delta A\|\|x_0\|}{\|A\|\|x_\ell\| + \|b\|} \leq \frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\| \|x_0\|}{\|A\| \|x_\ell\|}$$

concludes the proof. □

## 2.3 Numerical experiments

In this section we report on numerical experiments performed with MATLAB and machine precision  $\psi \sim 2.10^{-16}$ . They illustrate that the theoretical results established in exact arithmetic in the previous section do ensure the convergence of the relaxed GMRES even in the presence of round-off errors. For those experiments we only consider the strategies related to the convergence with respect to  $\eta_{A,b}(x_k)$ . Consequently the stopping criterion implemented by the algorithm is defined by (24). For those experiments we set  $\varepsilon_A$  and  $\varepsilon_b$  to zero. The matrices  $E_k$  are obtained using random matrices (MATLAB function `rand`) multiplied by a relevant scalar to match the upper-bound of the corresponding strategies. We set  $c = 1/4$ ,  $\varepsilon_c = \frac{1}{4\kappa(A)} \frac{4\kappa(A) - 1}{4\kappa(A) + 1}$  and  $\varepsilon > \varepsilon_c$ . We therefore consider the two strategies defined as follows

1. Strategy  $S^*$

$$\|E_k\| = \frac{\sigma_{\min}(A)}{4n} \min \left( 1, \frac{3\gamma^*}{\|\tilde{r}_{k-1}\|} \varepsilon_c \right) \text{ and } \|E_0\| = 0.$$

2. Strategy  $S^b$

$$\|E_k\| = \frac{\sigma_{\min}(A)}{4n} \min \left( 1, \frac{3\gamma^b}{\|\tilde{r}_{k-1}\|} \varepsilon_c \right) \text{ and } \|E_0\| = 0.$$

Note that  $S^b$  is derived from Remark 1. The strategy  $S^b$  is interesting, because it does not require the knowledge of  $\|x^*\|$ , but it only enables smaller perturbations than  $S^*$ . Finally, for the sake of completeness we depict in Algorithm 1 the details of the implementation of the relaxed GMRES algorithm.

---

**Algorithm 1** Relaxed GMRES with strategy  $S$

---

- 1: Choose a convergence threshold  $\varepsilon = \varepsilon_c + \varepsilon_g$
  - 2: Choose an initial guess  $x_0$
  - 3:  $r_0 = b - Ax_0$ ;  $\beta = \|r_0\|$
  - 4:  $v_1 = r_0/\|r_0\|$ ;
  - 5: **for**  $k = 1, 2, \dots$  **do**
  - 6:  $z = (A + E_k)v_k$ ,  $E_k$  being such that strategy  $S$  holds
  - 7: **for**  $i = 1$  **to**  $k$  **do**
  - 8:  $h_{i,k} = v_i^T z$
  - 9:  $z = z - h_{i,k}v_i$
  - 10: **end for**
  - 11:  $h_{k+1,k} = \|z\|$
  - 12:  $v_{k+1} = z/h_{k+1,k}$
  - 13: Solve the least-squares problem  $\min \|\beta e_1 - \bar{H}_k y\|$  for  $y_k$
  - 14: **if**  $\|\tilde{r}_k\| = \|\beta e_1 - \bar{H}_k y_k\| \leq \varepsilon_c \|A\| \|x_k\|$  **then**
  - 15: Set  $x_k = x_0 + V_k y_k$
  - 16: Exit
  - 17: **end if**
  - 18: **end for**
- 

Typical behaviours are presented for the linear system  $Ax = b$  in Figure 1 where  $A$  is the matrix PDE225 preconditioned using the incomplete LU factorization (i.e. ILU(t) [13]) with threshold  $t = 10^{-1}$  (i.e.  $A = U^{-1}L^{-1}A_{PDE225}$ ). The right-hand side  $b$  is such that  $x = (1, \dots, 1)^T$  is the solution of  $Ax = b$ . In this figure, we plot the convergence history that is, the backward errors  $\eta_{A,b}(x_k)$  along the iterations. The line without any tip is the convergence curve with the exact full GMRES, the line with  $\diamond$  (resp.  $+$ ) is the convergence of relaxed GMRES with strategy  $S^*$  (resp.  $S^b$ ). Notice that the three curves perfectly overlap. The dashed horizontal line corresponds to the targeted backward error  $\varepsilon = 10^{-13}$  and the line with  $\triangle$  (resp.  $*$ ) is the relative norm of the perturbation

$\frac{\|E_k\|}{\|A\|}$  associated with strategy  $S^*$  (resp.  $S^b$ ). It can be seen that as relaxed GMRES converges the accuracy of the matrix-vector product is significantly relaxed without preventing relaxed GMRES from converging to the targeted backward error. We also see that 4 extra relaxed GMRES iterations are performed before the convergence is detected while  $\eta_{A,b}(x_k)$  is already below  $\varepsilon$ . This delay is due to the stopping criterion implemented in the inexact algorithm, that relies on an upper-bound for  $\eta_{A,b}$  based on (5). If we had computed the exact residual associated to  $A$  along the iterations (which does not make much sense in the inexact framework), all the three strategies would have stopped at the same iteration.

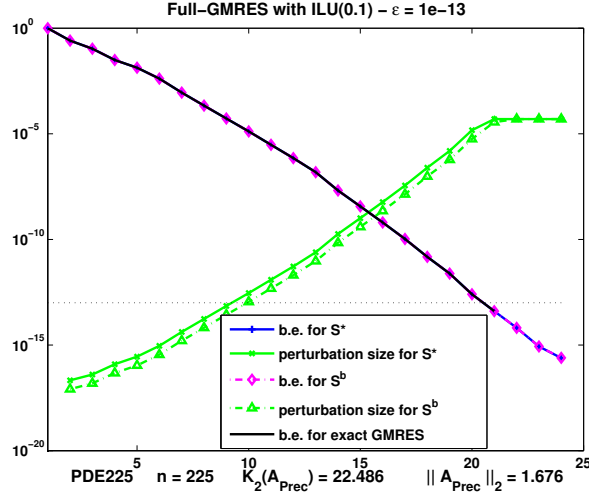


Figure 1: Relaxed GMRES with strategy  $S^*$  and  $S^b$  - PDE225 -  $\varepsilon = 10^{-13}$ .

In Figure 2, we display the same results for the matrix UTM 300 preconditioned by ILU( $10^{-3}$ ). Again it can be seen that the accuracy of the matrix-vector product can be relaxed without preventing the convergence to occur. However in that sort of extreme case, the perturbation size is smaller than  $\psi$  in the first iterations. This does not make much sense in finite precision calculation because the perturbation below  $\psi$  cannot be represented in finite precision. Finally, the example with matrix UTM 300 better illustrates that  $S^b$  is more conservative than  $S^*$ .

## 2.4 Design and behaviour of some relaxation heuristics

It is established in [6] that the classical (unperturbed) GMRES algorithm implemented using reliable orthogonal transformation is a backward stable method in finite precision. This means that the quantity  $\eta_{A,b}(x_k)$  is of the order of machine precision for some step  $\ell$ ,  $\ell \leq n$ . Let us use this result for a machine precision  $\psi = \varepsilon$ . Provided that all the operations occurring in the GMRES algorithm are

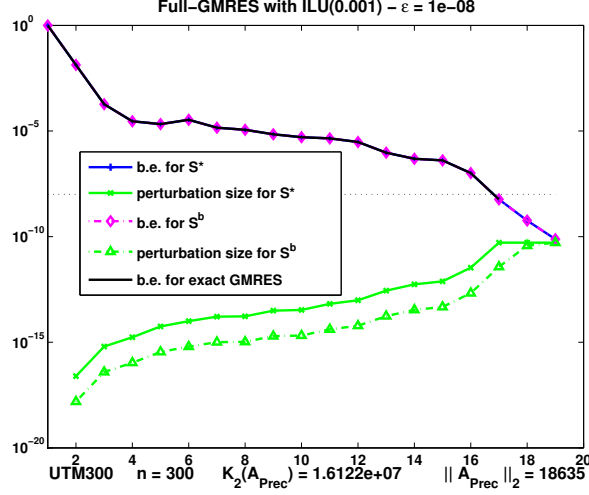


Figure 2: Relaxed GMRES with strategy  $S^*$  and  $S^b$  - UTM300 -  $\varepsilon = 10^{-8}$ .

performed with machine precision  $\varepsilon$ , a backward error  $\eta_{A,b}(x_k) \sim C\varepsilon$  can be reached, where  $C$  depends on the problem size and on the details of the arithmetic. In that context we can define three heuristics that are closely related to the strategies studied in the previous section. They are simply derived by thresholding the norm of the perturbations and prevent them from becoming smaller than  $\varepsilon$ .

This leads to the definition of the following three heuristics:

1. Heuristic  $S^*(\varepsilon)$

$$\|E_k\| = \max\left(\varepsilon\|A\|, \frac{\sigma_{\min}(A)}{4n} \min\left(1, \frac{3\gamma^*}{\|\tilde{r}_{k-1}\|} \varepsilon_c\right)\right),$$

and  $\|E_0\| = 0$ .

2. Heuristic  $S^b(\varepsilon)$

$$\|E_k\| = \max\left(\varepsilon\|A\|, \frac{\sigma_{\min}(A)}{4n} \min\left(1, \frac{3\gamma^b}{\|\tilde{r}_{k-1}\|} \varepsilon_c\right)\right),$$

and  $\|E_0\| = 0$ .

3. Heuristic  $S(\varepsilon)$

$$\|E_k\| = \varepsilon\|A\|.$$

This latter heuristic is related to exact GMRES run in a floating point arithmetic with machine precision  $\varepsilon$ .

In Table 1 we report on the number of iterations of exact GMRES and relaxed GMRES with the three heuristics on a set of matrices from Matrix Market. In that table we display the size of the matrix  $n$ , the threshold used for the  $ILU(t)$  preconditioner (‘-’ means no preconditioner), the targeted accuracy  $\varepsilon$  for  $\eta_{A,b}$ , the number of exact GMRES iterations is  $N_{\text{ex}}$ , the number of relaxed GMRES iterations with heuristic  $S(\varepsilon)$  ( $S^*(\varepsilon)$  and  $S^b(\varepsilon)$ ) is  $N_\varepsilon$  (resp.  $N_\varepsilon^*$  and  $N_\varepsilon^b$ ). For the inexact GMRES the stopping criterion is defined by (24). The right-hand side  $b$  is such that  $x^* = (1, \dots, 1)^T$  is the solution of  $Ax = b$ , where  $A$  stands again for the left preconditioned matrix. We first see that the three variants of inexact GMRES always converge to the targeted accuracy. Even though not displayed for those examples the convergence histories with the three heuristics perfectly overlap the one of exact GMRES along the iterations for all the matrices. We observed this behaviour for all the right-hand sides we have considered, not only for  $b = A(1, \dots, 1)^T$ . Thus the reason why the number of iterations for  $S(\varepsilon)$  is smaller than for  $S^*(\varepsilon)$  and  $S^b(\varepsilon)$  solely resides in the stopping criterion which induces a short delay in the convergence detection.

matrix	$n$	$t$	$\varepsilon$	Heuristics			
				$N_{\text{ex}}$	$N_\varepsilon$	$N_\varepsilon^*$	$N_\varepsilon^b$
e05r0400	236	$10^{-3}$	$10^{-14}$	21	21	23	23
e05r0000	236	$10^{-2}$	$10^{-06}$	25	25	31	31
GRE115	115	$10^{-1}$	$10^{-10}$	15	15	17	17
GRE185	185	$10^{-2}$	$10^{-14}$	21	21	22	22
GRE343	343	$10^{-1}$	$10^{-10}$	29	29	34	34
CAVITY03	317	$10^{-3}$	$10^{-10}$	18	18	21	21
PDE225	225	$10^{-1}$	$10^{-13}$	21	21	24	24
SAYLR1	238	$10^{-1}$	$10^{-11}$	29	29	32	32
UTM300	300	$10^{-3}$	$10^{-08}$	17	17	19	19
WEST0381	381	$10^{-2}$	$10^{-06}$	12	12	15	15
BFW398A	398	$10^{-1}$	$10^{-08}$	40	40	44	44

Table 1: # iterations of GMRES with various strategies.

The only example we have encountered that behaves differently is the **GRCAR** matrix with  $b = e_1$  and  $\varepsilon \geq 10^{-8}$  (also considered in [15]); in Figure 3 we display for that example the convergence history of  $\eta_{A,b}(x_k)$ . We see in this peculiar case that exact GMRES converges quickly; that is, the solution lies in a low dimensional Krylov subspace. If inexact matrix-vector products are used, this low dimensional invariant space is not captured as quickly and the convergence is significantly delayed. We mention that this behaviour disappears if a smaller value of  $\varepsilon$  (a larger Krylov space is required) is selected but still exists for larger  $\varepsilon$  (even smaller “invariant” space). It also disappears for other right-hand sides such as  $b = A(1, \dots, 1)^T$ .

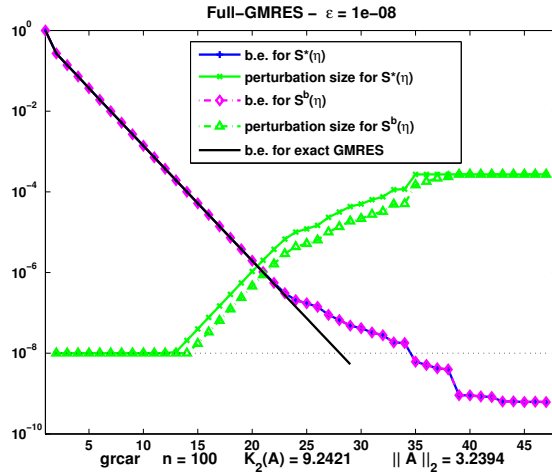


Figure 3: Convergence history of the inexact GMRES on GRCAR with  $b = e_1$

### 3 GMRES with relaxed right preconditioning

In practice GMRES is very often used with a preconditioner. When the preconditioner is used on the right, we solve the linear system  $AM^{-1}u = b$  and we recover the solution by  $x^* = M^{-1}u^*$ . An important class of preconditioners, often referred to as implicit preconditioners, tries to approximate the matrix  $A$  and then solves the linear system associated with  $M$  ( $M$  is chosen such that this operation is much easier with  $M$  than with  $A$ ). We are interested in the case where the application of the preconditioner is inexact: the system  $Mz = v$  involved in the preconditioned GMRES is not solved exactly. This happens for instance when block preconditioners are used. Examples are Schwarz preconditioners in domain decomposition [17] or more simply block Jacobi preconditioner where each block is solved approximately using an iterative scheme. In this case, we assume that the matrix-vector product by  $A$  is performed exactly but we have  $Mz = v + p$  where  $p$  is the residual associated with the preconditioning operation. Our aim is to find a strategy to monitor  $\|p\|$  in such a way that the convergence of GMRES reaches a backward error  $\varepsilon$ . A related study is developed in [15] where the focus is on the residual norm and the issue related to the recovery of  $x_\ell$  from  $u_\ell$  by the preconditioning  $Mx_\ell = u_\ell + p_\ell$  is not addressed.

The inexact Arnoldi relation now reads

$$AM^{-1}V_k + AM^{-1}P_k = (AM^{-1} + AM^{-1}P_kV_k^T)V_k = V_{k+1}\bar{H}_k, \quad (26)$$

where  $P_k = [p_1, \dots, p_k]$ ,  $p_k$  being the residual associated with the  $k$ -th preconditioning operation  $Mz_k = v_k + p_k$ . The following theorem shows a possible control of  $\|p_k\|$  so that the convergence of the relaxed right preconditioned GMRES is ensured. This result is closely related to Theorem 3.

**Theorem 5** Let  $\varepsilon_g$  be any positive real number. Let  $c$  and  $\varepsilon_c$  be such that  $0 < c < 1/2$ ,

$$\varepsilon_c \leq \frac{1-2c}{2\kappa(AM^{-1})} \frac{\kappa(AM^{-1})-c}{\kappa(AM^{-1})+c}.$$

Suppose that for all  $k$

$$\|p_k\| \leq \frac{1}{n\kappa(AM^{-1})} \min\left(c, \frac{(1-c)\gamma^p}{\|\tilde{r}_{k-1}\|} \varepsilon_g\right) \quad (27)$$

where

$\gamma^p = \frac{1-2\omega}{1-\omega} \|AM^{-1}\| \|u^*\| + \|b\|$  and  $\omega = c + \varepsilon_c \kappa(AM^{-1}) \frac{\kappa(AM^{-1})+c}{\kappa(AM^{-1})-c}$ . The exists  $\ell$ ,  $0 \leq \ell \leq n$ , such that the following stopping criterion is satisfied

$$\frac{\|\tilde{r}_\ell\|}{\|AM^{-1}\| \|u_\ell\|} \leq \varepsilon_c \quad \text{and} \quad \eta_{AM^{-1},b}(u_\ell) = \frac{\|r_\ell\|}{\|AM^{-1}\| \|u_\ell\| + \|b\|} \leq \varepsilon_u = \varepsilon_c + \varepsilon_g.$$

Proof: Setting  $E_k = AM^{-1}p_k v_k^T$ , the inexact Arnoldi relation (26) is exactly (3). The relaxation strategy of Theorem 3 writes

$$\|AM^{-1}p_k\| \leq \min\left(c \frac{\sigma_{\min}(AM^{-1})}{n}, \frac{(1-c)\sigma_{\min}(AM^{-1})}{n} \frac{\gamma^p}{\|\tilde{r}_{k-1}\|} \varepsilon_g\right).$$

The conclusion follows from  $\|AM^{-1}p_k\| \leq \|AM^{-1}\| \|p_k\|$ .

□

The result of Theorem 5 only gives a sufficient condition such that backward error associated with the approximated solution  $u_\ell$  of the *preconditioned* system  $AM^{-1}u = b$  is smaller than  $\varepsilon$ . To get the solution to the original system  $Ax = b$  an additional preconditioning operation  $Mx_\ell = t_\ell + p$  has to be performed, where  $p$  denotes the associated residual and  $\rho = \|p\|$ . From (1) follows that if the strategy of Theorem 5 is applied, the solution  $u_\ell$  obtained is such that there exists an  $n \times n$  matrix  $\Delta$  and an  $n$  vector  $\delta$  such that

$$(AM^{-1} + \Delta)u_\ell = b + \delta, \quad \text{with} \quad \max\left(\frac{\|\Delta\|}{\|AM^{-1}\|}, \frac{\|\delta\|}{\|b\|}\right) \leq \varepsilon_u$$

and from  $Mx_\ell = u_\ell + p$  follows that

$$\|Ax_\ell - b\| \leq \|\delta - \Delta u_\ell + AM^{-1}p\| \leq \|b\| \frac{\|\delta\|}{\|b\|} + \|AM^{-1}\| \frac{\|\Delta\|}{\|AM^{-1}\|} \|u_\ell\| + \|AM^{-1}\| \|p\|. \quad (28)$$

Using [9, Theorem 7.2] on  $u_\ell = (AM^{-1} + \Delta)^{-1}(b + \delta)$  proves that

$$\|u_\ell\| \leq \|u^*\| + \|MA^{-1}\| \frac{\|AM^{-1}\| \|u^*\| + \|b\|}{1 - \kappa(AM^{-1})\varepsilon_u} \varepsilon_u, \quad (29)$$

provided that  $\kappa(AM^{-1})\varepsilon_u < 1$ . Using (28) and (29) yields

$$\begin{aligned} \|Ax_\ell - b\| &\leq (\|b\| + \|AM^{-1}\| \|u^*\|)\varepsilon_u + \\ &\quad \frac{\|AM^{-1}\| \|u^*\| + \|b\|}{1 - \kappa(AM^{-1})\varepsilon_u} \kappa(AM^{-1})\varepsilon_u + \|AM^{-1}\| \|p\| \\ &\leq \frac{\|AM^{-1}\| \|u^*\| + \|b\|}{1 - \kappa(AM^{-1})\varepsilon_u} \varepsilon_u + \|AM^{-1}\| \rho, \end{aligned}$$

which gives an upper bound for  $\eta_{A,b}(x_\ell)$  that depends on  $\rho$  and  $\varepsilon_u$ . We summarize this result in the next theorem.

**Theorem 6** *Suppose that the relaxed right-preconditioned GMRES is run on  $AM^{-1}u = b$  under the assumptions of Theorem 5 and that  $u_\ell$  is the corresponding estimate of  $u^*$ . Suppose in addition that  $Mx_\ell = u_\ell + p$ , with  $\rho = \|p\|$ , and that  $\kappa(AM^{-1})\varepsilon_u < 1$ . The backward error  $\eta_{A,b}(x_\ell)$  of  $x_\ell$  considered as a solution of  $Ax = b$  satisfies*

$$\eta_{A,b}(x_\ell) \leq \frac{1}{\|A\| \|x_\ell\| + \|b\|} \left( \frac{\|AM^{-1}\| \|u^*\| + \|b\|}{1 - \kappa(AM^{-1})\varepsilon_u} \varepsilon_u + \|AM^{-1}\| \rho \right).$$

We present a numerical illustration for Theorem 6 in order to demonstrate the effect of the last preconditioning operation on the backward error of the computed solution  $x_\ell$ . We consider the matrix E05R0000 with an  $ILLU(10^{-2})$  as right preconditioner and a targeted backward error on the preconditioned system  $\eta_{AM^{-1},b}(u_\ell) \leq 10^{-10}$ . To recover the solution we perform a final preconditioning step  $Mx_\ell = u_\ell + p$ , where a random vector  $p$  is chosen such that  $\|p\|$  is equal to a prescribed quantity  $\rho$ . We report in Table 2 the backward errors  $\eta_{A,b}(x_\ell)$  obtained at the end of the process (inexact right preconditioned GMRES and the final preconditioned step  $Mx_\ell = u_\ell + p$ ) and the associated upper-bound presented in Theorem 6. We see that the upper-bound is tight on this example, and that the backward error on  $x_\ell$  strongly depends on the accuracy of the final  $Mx_\ell = u_\ell + p$  as monitored by the quantity  $\rho = \|p\|$ .

$\rho$	$\eta_{A,b}(x_\ell)$	upper-bound
$10^{-14}$	$4 \cdot 10^{-14}$	$6 \cdot 10^{-14}$
$10^{-11}$	$3 \cdot 10^{-12}$	$1 \cdot 10^{-11}$
$10^{-08}$	$3 \cdot 10^{-09}$	$1 \cdot 10^{-08}$
$10^{-05}$	$3 \cdot 10^{-06}$	$1 \cdot 10^{-05}$
$10^{-01}$	$7 \cdot 10^{-04}$	$3 \cdot 10^{-03}$

Table 2: GMRES with relaxed right preconditioner.

**Remark 2** *We mention that similar results can be derived for inexact left preconditioning. In that context, the quantity  $\varepsilon_b$  from Theorem 4 enables us to account for the inaccuracy in the first preconditioning step.*

**Remark 3** We consider FGMRES with a varying preconditioner satisfying  $Mz_k = v + p_k$ , where  $p_k$  is the residual vector which norm is used to control the convergence of the algorithm. FGMRES is very close to the inexact right preconditioned GMRES in the sense that the same Arnoldi relation (26) holds, which can be rewritten  $AZ_k = V_{k+1}\bar{H}_k$ . Contrary to the inexact GMRES framework, [13, Proposition 9.3] shows that when a breakdown occurs in FGMRES and when the associated  $H_\ell$  is nonsingular, not only  $\tilde{r}_\ell$  is zero but also  $x_\ell = Z_\ell y_\ell = A^{-1}b$ . In other words the nonsingularity of  $H_\ell$  yields that at the breakdown,  $x_\ell$  is the solution to  $Ax = b$ . Theorems 1 and (27) show that the nonsingularity of  $H_\ell$  is guaranteed if  $\|p_k\| \leq \frac{c}{n\kappa(AM^{-1})}$ .

Theorem 6 mainly emphasizes the role played by the last preconditioning step. Its use in real applications would deserve additional efforts to find reasonable estimates of the quantities involved in the control such that  $\kappa(AM^{-1})$  or  $\|AM^{-1}\|$ .

## 4 Some remarks on inexact restarted GMRES

In this paper we are interested in relaxation techniques for which theoretical convergence of GMRES can be established. This is the main reason why we have only considered strategies in the context of full GMRES. There is a clear interest in practice to use the restarted GMRES(m) algorithm [14]. Unfortunately, no general theoretical result exists on the convergence of restarted GMRES. Consequently it is hopeless to establish any theoretical result for relaxed GMRES(m) at the date of today even though numerical experiments [3, 7, 16] show that the convergence might be observed in practice. In that section, we consider a variant of restarted GMRES where the restart is not governed by the dimension of the Krylov space affordable in term of memory but rather by a targeted backward error  $\eta$ . We study this variant of restarted GMRES as a fixed point iteration scheme, where at each restart (i.e. each fixed point iteration) the residual can be inaccurately computed. This fixed point iteration is summarized in Algorithm 2. We notice that step 4 can be performed by any linear solver, if we consider full-GMRES we end-up with what we call GMRES( $\eta$ ). Finally, if step 4 is replaced by  $m$  steps of full-GMRES, Algorithm 2 reduces to classical GMRES(m) with inexact initial residual.

---

### Algorithm 2 Fixed point scheme

---

- 1: Choose an initial guess  $x_0$
- 2: **for**  $k = 1, 2, \dots$  **do**
- 3:   Compute  $r_k = (b + f_k) - (A + E_k)x_k$
- 4:    $s_k = \text{solve}(A, r_k, \eta)$
- 5:    $x_{k+1} = x_k + s_k$
- 6: **end for**

where  $s_k$  is such that  $\eta_{A,b}(s_k) \leq \eta$  with  $\|E_k\| \leq \varepsilon_A \|A\|$  and  $\|f_k\| \leq \varepsilon_b \|b\|$ .

---

**Theorem 7** Assuming that  $\varepsilon_A \kappa(A) < 1/5$  and that  $\eta \kappa(A) < 1/5$ , for  $\ell$  large enough

$$\frac{\|x_\ell - x^*\|}{\|x^*\|} \leq 10\kappa(A, b) \max(\varepsilon_A, \varepsilon_b)$$

where  $x_\ell$  is obtained using Algorithm 2. In particular, this holds for  $s_k$  computed with relaxed GMRES of Section 2.2.

Proof: Let consider the fixed-point iteration

$$x_{k+1} = x_k + u_k$$

and denote  $\delta_k = x_k - x^*$  where  $u_k = \text{solve}(A, \tilde{r}_k, \eta)$  with  $\tilde{r}_k = b - \Delta b_k^{res} - (A + \Delta A_k^{res})x_k$  and  $u_k$  is such that it exist  $\Delta A_k^{sol}$  and  $\Delta b_k^{sol}$  such that

$$\|\Delta A_k^{sol}\| \leq \eta \|A\| \quad , \quad \|\Delta r_k\| \leq \eta \|b - Ax_k\| \leq \|A\| \|\delta_k\| \quad (30)$$

and

$$\|\Delta A_k^{res}\| \leq \varepsilon_A \|A\| \quad , \quad \|\Delta b_k^{res}\| \leq \varepsilon_b \|b\|$$

We have  $x_{k+1} = x_k + (A + \Delta A_k^{sol})^{-1}(\tilde{r}_k + \Delta r_k)$  which gives

$$\begin{aligned} e_{k+1} &= \delta_k + (A + \Delta A_k^{sol})^{-1} (-A\delta_k - \Delta b_k^{res} - \Delta r_k - \Delta A_k^{res}(\delta_k + x^*)) \\ &= \left( I - (A + \Delta A_k^{sol})^{-1} (A + \Delta A_k^{res}) \right) \delta_k + \\ &\quad (A + \Delta A_k^{sol})^{-1} (\Delta b_k^{res} - \Delta A_k^{res} x^* - \Delta r_k) \end{aligned}$$

and using (30)

$$\begin{aligned} \|e_{k+1}\| &\leq \frac{\|A^{-1}\| (\|\Delta A_k^{sol}\| + \|\Delta A_k^{res}\| + \eta) \|\delta_k\| + (\varepsilon_b \|b\| + \varepsilon_A \|A\| \|x^*\|)}{1 - \|A^{-1}\| \|\Delta A_k^{sol}\|} \\ &\leq \frac{\kappa(A)(2\eta + \varepsilon_A) \|\delta_k\| + \max(\varepsilon_b, \varepsilon_A) \|A^{-1}\| (\|b\| + \|A\| \|x^*\|)}{1 - \eta \kappa(A)}. \end{aligned}$$

Under the assumption  $\frac{\kappa(A)(2\eta + \varepsilon_A)}{1 - \eta \kappa(A)} < 1$ , that is for instance  $\eta \kappa(A) < 1/5$  and  $\varepsilon_A \kappa(A) < 1/5$ , we have  $\limsup_{k \rightarrow \infty} \|e_k\| \leq 5 \max(\varepsilon_b, \varepsilon_A) \|A^{-1}\| (\|b\| + \|A\| \|x^*\|)$  which concludes the proof.  $\square$

The above theorem shows that for a large enough  $\ell$  the forward error is bounded by ten times the product of a backward error by the normwise condition number of the linear system [9, p. 121] that is defined by  $\kappa(A, b) = \frac{\|A^{-1}\|}{\|x^*\|} (\|A\| \|x^*\| + \|b\|)$  where  $x^* = A^{-1}b$ . This is reasonable from a perturbation theory point of view.

In Figure 4 we display the convergence history of GMRES( $\eta$ ) for two different values of  $\varepsilon_A$  and  $\varepsilon_b$  and step 4 is implemented by relaxed GMRES. The curves with  $\times$  represents the norm of the perturbations involved in the relaxed GMRES. It can be seen that as long as the initial residual of GMRES( $\eta$ ) is performed with perturbation of the size of  $\varepsilon$ , the matrix-vector product involved in the iterations can be performed with much less accuracy.

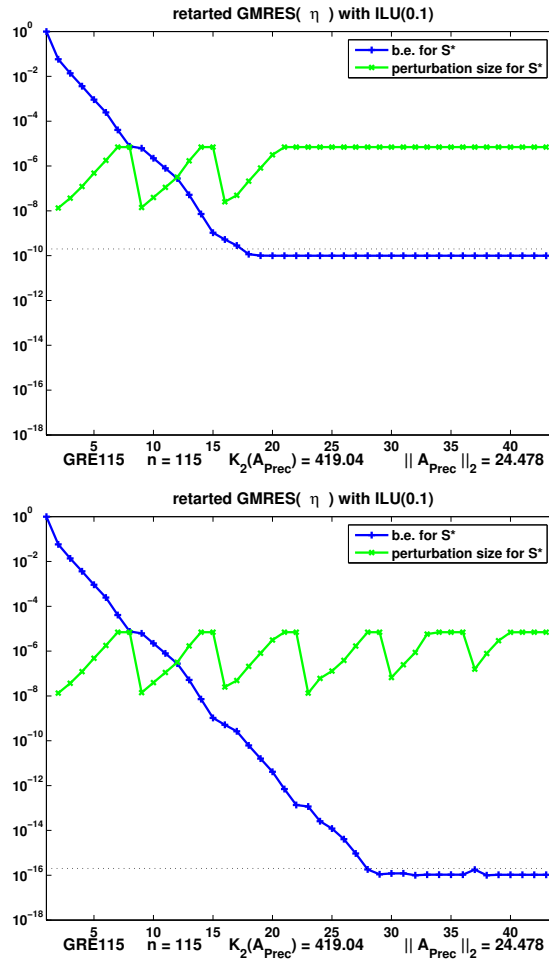


Figure 4: Restarted GMRES( $\varepsilon$ ) with inexact matrix-vector product - GRE115 - with  $\varepsilon_A = \varepsilon_b = 10^{-10}$  (top) and  $\varepsilon_A = \varepsilon_b = 10^{-16}$  (bottom)

## 5 Conclusion

In this paper we are interested either in relaxation techniques for which theoretical convergence of relaxed GMRES can be established or in heuristics that are closely related to the backward stability of the GMRES algorithm with reliable orthogonalization schemes. The proposed strategies ensure the convergence in backward error  $\eta_{A,b}$  or  $\eta_b$  down to a prescribed accuracy  $\varepsilon$ . Finally, our strategies rely on the knowledge of  $\sigma_{\min}(A)$ , the solution norm that might be difficult to estimate. Further research is needed to establish similar convergence results while replacing these quantities by others that are simpler to estimate.

## Acknowledgments

We would like to thank the anonymous referees for their valuable remarks, questions and comments that enable us to substantially improve this paper.

## References

- [1] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1995.
- [2] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. Van der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM, Philadelphia, PA, second edition, 1994.
- [3] A. Bouras and V. Frayssé. Inexact matrix-vector products in Krylov methods for solving linear systems: a relaxation strategy. *SIAM Journal on Matrix Analysis and Applications*, 2005. To appear.
- [4] A. Bouras, V. Frayssé, and L. Giraud. A relaxation strategy for inner-outer linear solvers in domain decomposition methods. Technical Report TR/PA/00/17, CERFACS, Toulouse, France, 2000.
- [5] N. Cundy, J. van den Eshof, A. Frommer, S. Krieg, Th. Lippert, and K. Schaefer. Numerical methods for the QCD overlap operator: III Nested iterations. *Computer Physics Communications*, 2005. To appear.
- [6] J. Drkošová, M. Rozložník, Z. Strakoš, and A. Greenbaum. Numerical stability of the GMRES method. *BIT*, 35:309–330, 1995.
- [7] L. Giraud, S. Gratton, and J. Langou. A note on relaxed and flexible GMRES. Technical Report TR/PA/04/41, CERFACS, Toulouse, France, 2004. Available on <http://www.cerfacs.fr/algor/reports>.
- [8] A. Greenbaum. *Iterative methods for solving linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997.

- [9] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2002.
- [10] G. L. G. Sleijpen J. van den Eshof and M. van Gijzen. Relaxation strategies for nested krylov methods. *Journal of Computational and Applied Mathematics*, 2005. To appear.
- [11] K. Mer-Nkonga and F. Collino. The fast multipole method applied to a mixed integral system for time-harmonic Maxwell’s equations. In B. Michielsen and F. Decavèle, editors, *European symposium on numerical methods in electromagnetics*, pages 121–126, 2002.
- [12] Y. Saad. A flexible inner-outer preconditioned GMRES algorithm. *SIAM Journal Scientific Computing*, 14:461–469, 1993.
- [13] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, 2003. Second edition.
- [14] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7:856–869, 1986.
- [15] V. Simoncini and D. B. Szyld. Theory of inexact Krylov subspace methods and applications to scientific computing. *SIAM Journal Scientific Computing*, 25:454–477, 2003.
- [16] G. L. G. Sleijpen, J. van den Eshof, and M. B. van Gijzen. Restarted GMRES with inexact matrix–vector products. Technical Report TR/PA/04/75, CERFACS, Toulouse, France, 2004.
- [17] B.F. Smith, P.E. Bjørstad, and W. Gropp. *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*. Cambridge University Press, 1996.
- [18] J. van den Eshof and G. L. G. Sleijpen. Inexact Krylov subspace methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 26(1):125–153, 2004.
- [19] J. S. Warsa, M. Benzi, T. A. Warein, and J. E. Morel. Preconditioning a mixed discontinuous finite element method for radiation diffusion. *Numerical Linear Algebra with Applications*, 11(8-9):795–811, 2004.