A Comparison of Some GMRES-like Methods

C. Vuik
Faculty of Technical Mathematics and Informatics
Delft University of Technology, The Netherlands

and

H. A. van der Vorst
Department of Mathematics
University Utrecht, The Netherlands

Submitted by Richard A. Brualdi

ABSTRACT

GMRES and CGS are well known iterative methods for the solution of certain sparse linear systems with a nonsymmetric matrix. These methods have been compared experimentally in many studies, and specific observations on their convergence behavior have been reported. A new iterative method to solve a nonsymmetric system has been proposed by Eirola and Nevanlinna. The purpose of this paper is to investigate this method and to compare it with GMRES. We have seen problems for which this method is more efficient than GMRES. The original method has as drawbacks that it is not scaling invariant and that it may suffer from numerical instability, but it is shown that these deficiencies can be repaired. A method proposed by Broyden (1969) seems related to the new method and is therefore included in the comparison.

INTRODUCTION

In this paper we compare the GMRES method (Saad and Schultz, 1986), the EN method (Eirola and Nevanlinna, 1989), and the B method (Broyden, 1969). Our main motivation to study the EN method is that it deepens our insight into projection-type methods, which may lead to better iterative methods. Descriptions and some relevant properties of these methods are given in Section 1. In Section 2 we describe numerical experiments for EN, which motivate the theoretical analysis of Section 3. In that section we give a

655 Avenue of the Americas, New York, NY 10010 0024-3795/92/$3.50

relation between the EN and the GMRES method. Subsequently we compare the efficiency of the two methods. Though in some cases the EN method is more efficient than the GMRES method, this is not the case in general. In Section 4 we show that the convergence and the stability properties of EN are not scaling invariant, as they are for GMRES and other projection methods, and we also show how this can be rectified to the advantage of the EN method. Furthermore, we describe some problems for which EN diverges and GMRES converges. In Section 5 we consider a variant of the EN method, which is algebraically equivalent to the GMRES method. This enables us to make a better comparison between GMRES and EN, and it gives more insight into GMRES. Finally, in Section 6 we compare the EN method with the B method and a general class of methods given in Broyden (1970). Furthermore we compare the efficiency of B and GMRES. From these comparisons it appears that the most efficient and robust method is the implementation of the full GMRES method as described in, e.g., Saad and Schultz (1986) and Van der Vorst (1989). However, it appears from experiments that if the iterative methods (EN and GMRES) are restarted, then EN can be much more efficient than GMRES. This aspect is a subject for further study and is not reported on in this paper.

1. GMRES, EN, AND B METHODS

The GMRES method was originally proposed in Saad and Schultz (1986). We use results in Huang and Van der Vorst (1989) for understanding the convergence behavior of GMRES. Consider the linear system $Ax = b$ with $x, b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ nonsingular. The Krylov subspace $K^k(A; r_0)$ is defined by $K^k(A; r_0) = \text{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$. The kth iterate x_k is written as $x_k = x_0 + z_k$, where $z_k \in K^k(A; r_0)$ and $r_0 = b - Ax_0$. In the GMRES method the vector z_k is chosen as the vector which solves the linear least-squares problem

$$z_k = \arg \min_{z \in K^k(A; r_0)} \| b - A(x_0 + z) \|_2. \tag{1.1}$$

From this definition it follows that

$$\|r_k\|_2 = \min_{z \in K^k(A; r_0)} \|b - Ax_0 - Az\|_2 = \min_{\alpha_1, \ldots, \alpha_k \in \mathbb{R}} \left\| r_0 + \sum_{i=1}^{k} \alpha_i A^i r_0 \right\|_2 \tag{1.2}$$

In the EN method we take a different splitting of the matrix in each iteration step:

$$A = H_k^{-1} - R_k,$$

which leads to the basic iteration method

$$x_k = x_{k-1} + H_k r_{k-1}.$$

The key idea is to improve H_k from step to step by (cheap) rank-1 updates:

$$H_k = H_{k-1} + u_{k-1} v_{k-1}^T.$$

For the kth step this leads to

$$\begin{aligned} r_k &= r_{k-1} - A\big(H_{k-1} + u_{k-1} v_{k-1}^T\big) r_{k-1} \\ &= (I - AH_{k-1}) r_{k-1} - \mu_{k-1} A u_{k-1} \end{aligned}$$

with $\mu_{k-1} = v_{k-1}^T r_{k-1}$.

The ideal choice for u_{k-1} would have been such that

$$\mu_{k-1} A u_{k-1} = (I - AH_{k-1}) r_{k-1},$$

or

$$\mu_{k-1} u_{k-1} = A^{-1}(I - AH_{k-1}) r_{k-1}.$$

If H_{k-1}^{-1} is a suitable split-off part of A, then A^{-1} can be replaced by H_{k-1}, and this motivates the choice for u_{k-1}:

$$u_{k-1} = H_{k-1}(I - AH_{k-1}) r_{k-1}.$$

The choice for v_{k-1} now follows by minimizing $\|r_k\|_2$ as a function of μ_{k-1}:

$$\mu_{k-1} = \frac{(Au_{k-1})^T (I - AH_{k-1}) r_{k-1}}{\|Au_{k-1}\|_2^2},$$

so that

$$v_{k-1} = \frac{1}{\|Au_{k-1}\|_2^2} (I - AH_{k-1})^T A u_{k-1}$$

is an obvious choice.

This leads to the following algorithm (Eirola and Nevanlinna, 1989, pp. 512, 513):

1. Given x_0, H_0, compute r_0 and take $k = 0$.
2. $E_k = I - AH_k$, $u_k = H_k E_k r_k$, $v_k = E_k^T A u_k / \|Au_k\|_2^2$.
3. $H_{k+1} = H_k + u_k v_k^T$, $x_{k+1} = x_k + H_{k+1} r_k$, $r_{k+1} = b - Ax_{k+1}$.
4. Stop if $\|r_{k+1}\|_2$ is small enough; otherwise $k := k + 1$ and return to step 2.

The only difference between EN and GMRES is the choice of u_k. By taking $u_k = H_k r_k$ instead of $u_k = H_k E_k r_k$, we obtain an iterative method algebraically equivalent to GMRES.

The following equalities and definitions will be used in our analysis:

$$c_k \equiv Au_k / \|Au_k\|_2, \tag{1.3}$$

$$E_{k+1} = (I - P_k) E_0, \qquad P_k = \sum_{i=0}^{k} c_i c_i^T, \quad \text{and} \quad c_i^T c_j = 0 \quad \text{for } i \neq j, \tag{1.4}$$

$$r_{k+1} = E_{k+1} r_k. \tag{1.5}$$

Equation (1.4) only holds if all H_k are nonsingular. Therefore, in the case that H_{k+1} is singular whereas H_k is nonsingular we take $H_{k+1} = H_k$ (see Eirola and Nevanlinna, 1989, p. 518). The following property may be used to check whether H_{k+1} is singular.

Property 1.6 (Eirola and Nevanlinna, 1989, p. 518, Proposition 2.1). Assume H_k is nonsingular. Then H_{k+1} is singular if and only if $c_k^T E_0 r_k = 0$.

The description of the algorithm given above is suitable for analysis; however, in order to save computational work we prefer the following implementation given in Eirola and Nevanlinna (1989, p. 519): At step k, $x_k, r_k, u_0, \ldots, u_{k-1}, c_0, \ldots, c_{k-1}$ are known. Then compute

1. $\alpha_i = c_i^T (r_k - AH_0 r_k)$ for $i = 0, \ldots, k-1$, $\eta = H_0 r_k + \sum_{i=0}^{k-1} \alpha_i u_i$, $\xi = r_k - A\eta$;
2. $\beta_i = c_i^T (\xi - AH_0 \xi)$ for $i = 0, \ldots, k-1$, $u_k = \tau(H_0 \xi + \sum_{i=0}^{k-1} \beta_i u_i)$, $c_k = Au_k$, where τ is such that $\|c_k\|_2 = 1$;
3. $x_{k+1} = x_k + \eta + u_k c_k^T \xi$, $r_{k+1} = \xi - c_k c_k^T \xi$.

In the sequel, EN1 denotes the given implementation.

In another implementation given in Eirola and Nevanlinna (1989; p. 519), ξ and c_k are computed as follows: $\xi = r_k - AH_0 r_k - \sum_{i=0}^{k-1} \alpha_i c_i$ and $c_k =$

$\tau(AH_0\xi + \sum_{i=0}^{k-1}\beta_i c_i)$. This implementation is used in situations where it is more efficient to compute a linear combination of $k+1$ vectors instead of multiplying one vector by A. Note that ξ is the component of $r_k - AH_0 r_k$ orthogonal to $\text{span}\{c_0, \ldots, c_{k-1}\}$. Hence β_i is equal to $-c_i^T AH_0\xi$, which implies that c_k is the normalized component of $AH_0\xi$ orthogonal to $\text{span}\{c_0, \ldots, c_{k-1}\}$. In this implementation the vectors ξ and c_k are made orthogonal by the Gram-Schmidt process. For stability reasons we propose the following implementation (EN2) based on the modified Gram-Schmidt process:

1. $\xi^{(0)} = (I - AH_0)r_k$, $\eta^{(0)} = H_0 r_k$, $\alpha_i = c_i^T \xi^{(i)}$, $\xi^{(i+1)} = \xi^{(i)} - \alpha_i c_i$, $\eta^{(i+1)} = \eta^{(i)} + \alpha_i u_i$, $i = 0, \ldots, k-1$;
2. $c_k^{(0)} = AH_0\xi^{(k)}$, $u_k^{(0)} = H_0\xi^{(k)}$, $\beta_i = -c_i^T c_k^{(i)}$, $c_k^{(i+1)} = c_k^{(i)} + \beta_i c_i$, $u_k^{(i+1)} = u_k^{(i)} + \beta_i u_i$, $i = 0, \ldots, k-1$, $c_k = c_k^{(k)} / \|c_k^{(k)}\|_2$, $u_k = u_k^{(k)} / \|c_k^{(k)}\|_2$;
3. $x_{k+1} = x_k + \eta^{(k)} + u_k c_k^T \xi^{(k)}$, $r_{k+1} = (1 - c_k c_k^T)\xi^{(k)}$.

In our experiments the stability properties of EN1 and EN2 have appeared to be more or less equivalent.

In the B method, a nonsingular matrix $H_0 \in \mathbb{R}^{n\times n}$ must also be specified, which again is viewed as an approximation to the inverse of A. The algorithm runs as follows (Broyden, 1969, p. 94):

1. given x_0, H_0, compute r_0 and take $k = 0$,
2. $p_k = H_k r_k$, $x_{k+1} = x_k + p_k$, $r_{k+1} = b - Ax_{k+1}$,
3. $y_k = r_k - r_{k+1}$,
4. $H_{k+1} = H_k - (H_k y_k - p_k) p_k^T H_k / p_k^T H_k y_k$, $k := k+1$, and return to step 2.

2. NUMERICAL EXPERIMENTS

In order to get some idea of the convergence behavior of the EN method, we report on some numerical experiments. These experiments have been carried out in double precision arithmetic ($\approx$ is decimal places) on a HP9000-845 computer. Our test matrices and right-hand sides are taken from Huang and Van der Vorst (1989, pp. 16, 17). These matrices are of the form $A = SBS^{-1}$ with $A, S, B \in \mathbb{R}^{100\times 100}$. We have selected S to be equal to

$$S = \begin{bmatrix} 1 & \beta & & 0 \\ & 1 & \ddots & \\ & & \ddots & \beta \\ 0 & & & 1 \end{bmatrix}.$$

The system $Ax = b$ is solved for right-hand sides such that $x = (1, \ldots, 1)^T$ (experiments with other choices of x show more or less the same convergence behavior). In these experiments we take $H_0 = I$ and $x_0 = (0, \ldots, 0)^T$. The matrices in our testset are as follows [the numbering refers to the problems in Huang and Van der Vorst (1989, p. 17)]:

PROBLEM P6.

$$B = \begin{bmatrix} 1 & & & & & 0 \\ & 1+\alpha & & & & \\ & & 3 & & & \\ & & & 4 & & \\ & & & & \ddots & \\ 0 & & & & & 100 \end{bmatrix}$$

(double eigenvalue for $\alpha \to 0$).

PROBLEM P7.

$$B = \begin{bmatrix} 1 & \alpha & & & & 0 \\ -\alpha & 1 & & & & \\ & & 3 & & & \\ & & & 4 & & \\ & & & & \ddots & \\ 0 & & & & & 100 \end{bmatrix}$$

(conjugate eigenpair $1 \pm \alpha i$).

PROBLEM P8.

$$B = \begin{bmatrix} 1 & 1 & & & & & & 0 \\ -1 & 1 & & & & & & \\ & & 1+\alpha & 1+\alpha & & & & \\ & & -1-\alpha & 1+\alpha & & & & \\ & & & & 5 & & & \\ & & & & & 6 & & \\ & & & & & & \ddots & \\ 0 & & & & & & & 100 \end{bmatrix}$$

(two conjugate eigenpairs which come close for $\alpha \to 0$).

PROBLEM P9.

$$B = \begin{bmatrix} 1 & 1 & & & & 0 \\ & 1 & & & & \\ & & 3 & & & \\ & & & 4 & & \\ & & & & \ddots & \\ 0 & & & & & 100 \end{bmatrix}$$

(defect matrix with Jordan block of order 2).

For these problems we have plotted the convergence behavior of the EN method in terms of the reduction factors $\|r_{k+1}\|_2 / \|r_k\|_2$, for different values of α and β. In order to facilitate comparison, different curves have been plotted in the same figure. The lowest curve is always plotted on the right scale. Each successive curve has been raised by 0.1 vertically with respect to the previous one.

The results for $B := B/100$ (an explanation of this seemingly awkward choice is given in Section 4) are given in Figures 1, 3, 5, and 7. These figures are in a qualitative sense largely the same as Figures 2, 4, 6, and 8 for GMRES,

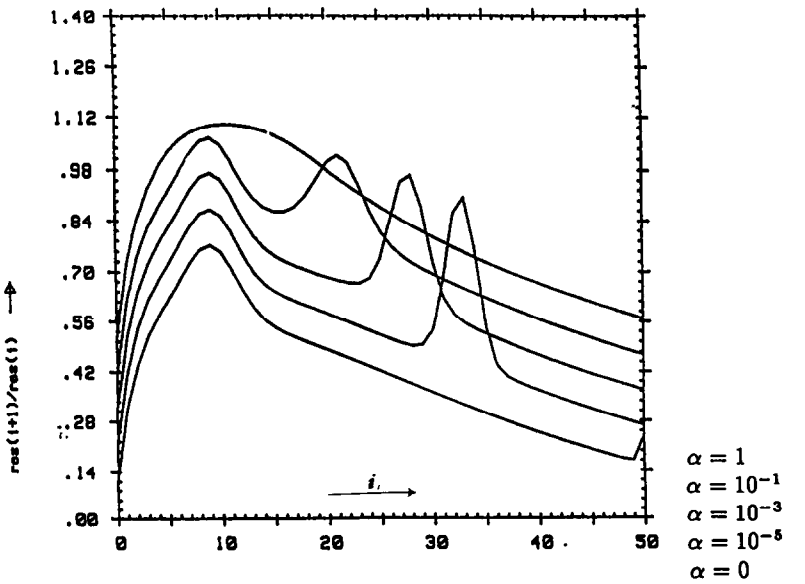

FIG. 1. Problem P6, $\beta = 0.9$, EN.

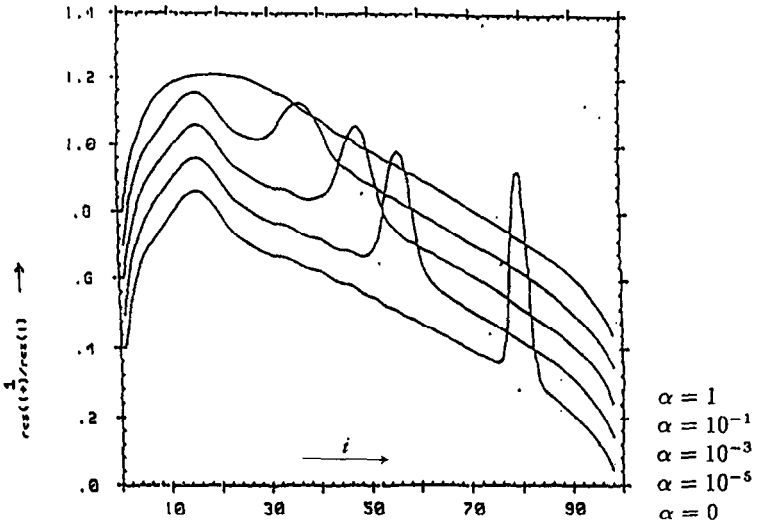

FIG. 2. Problem P6, $\beta = 0.9$, GMRES.

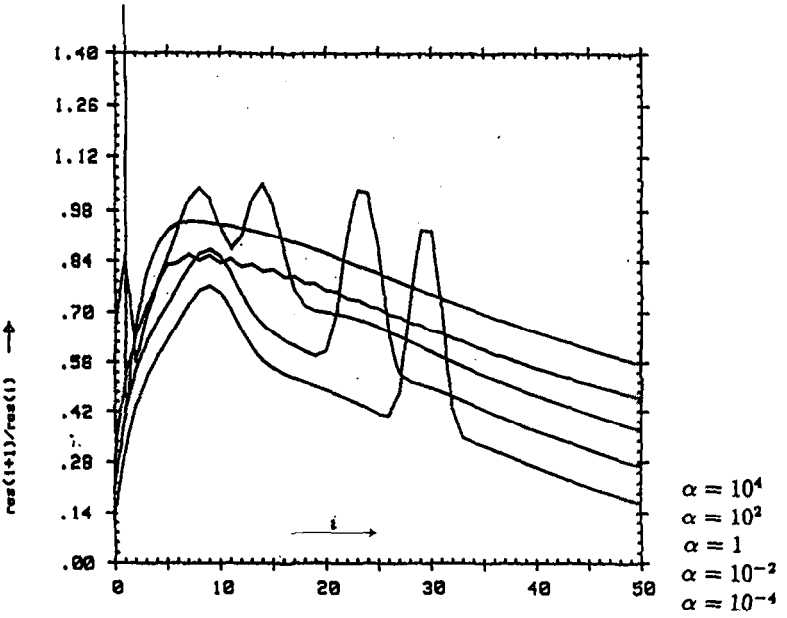

FIG. 3. Problem P7, $\beta = 0.9$, EN.

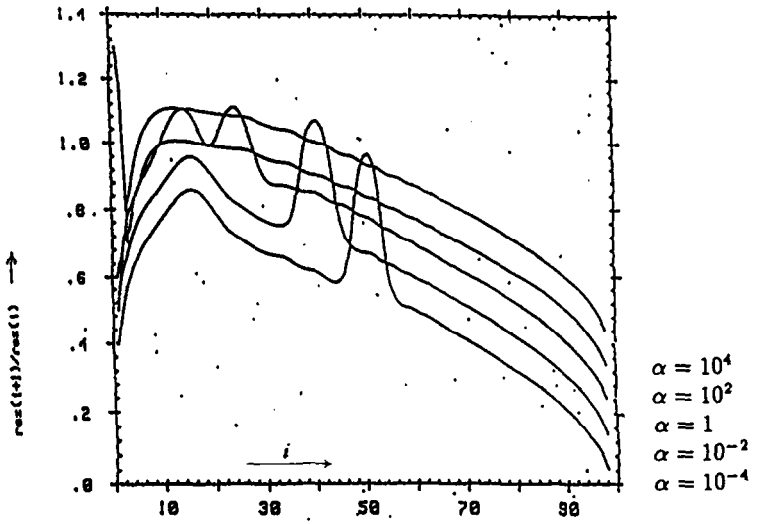

FIG. 4. Problem P7, $\beta = 0.9$, GMRES.

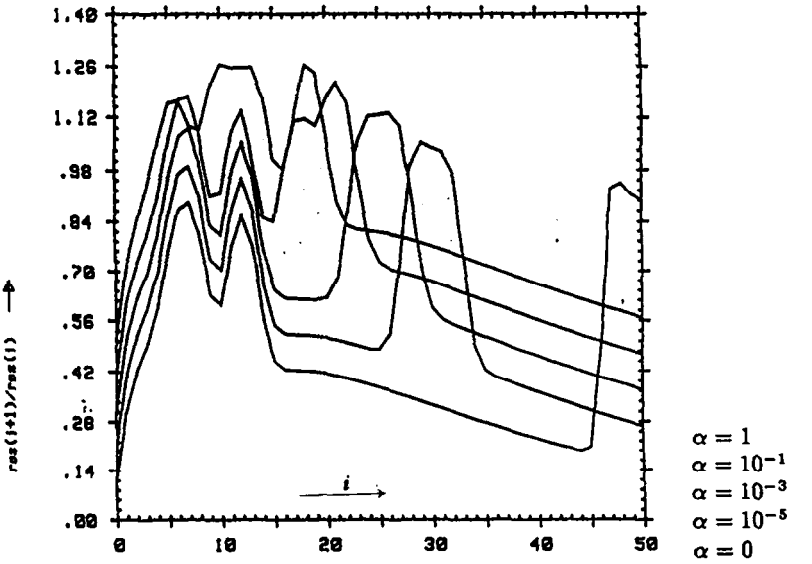

FIG. 5. Problem P8, $\beta = 0.9$, EN.

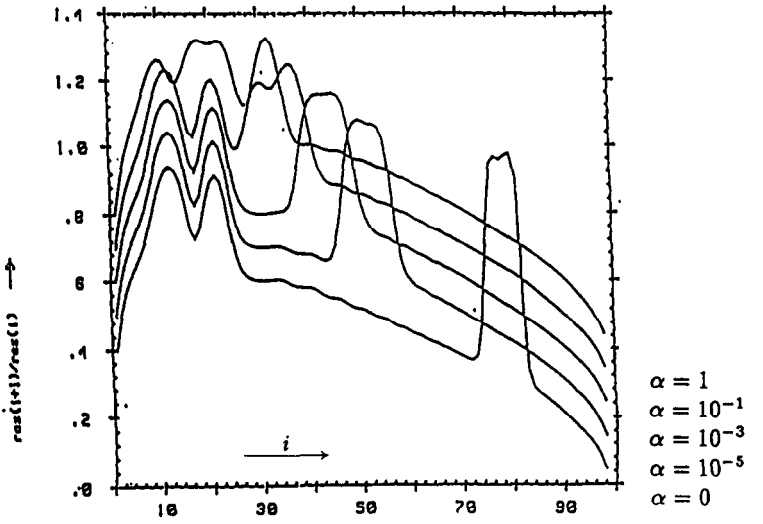

FIG. 6. Problem P8, $\beta = 0.9$, GMRES.

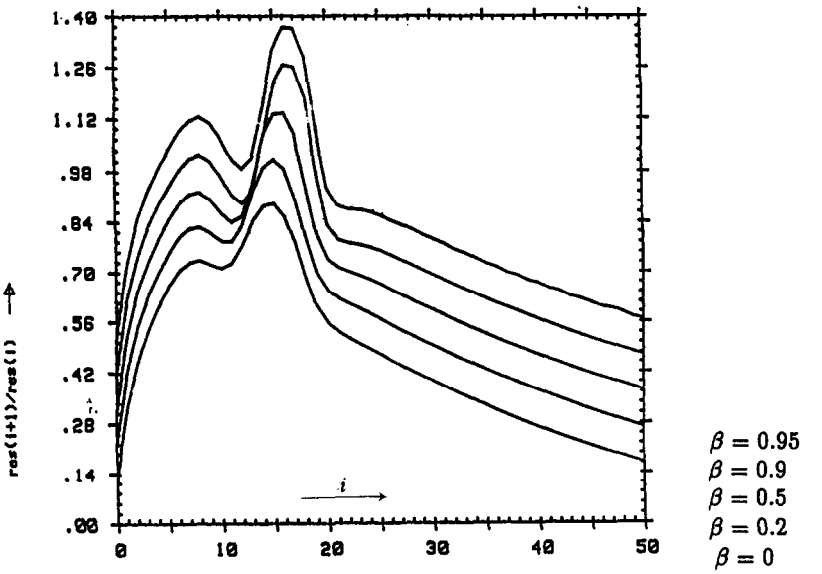

FIG. 7. Problem P9, EN.

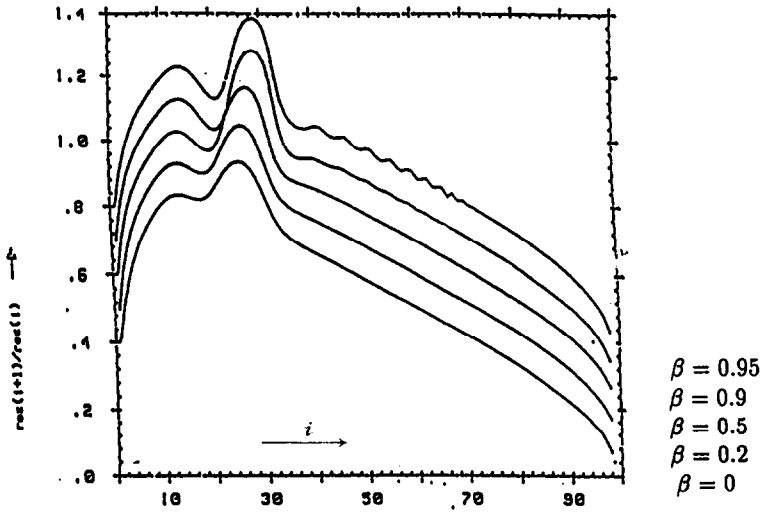

FIG. 8. Problem P9, GMRES.

obtained from Huang and Van der Vorst (1989, Figures 17, 18, 20, 21). This leads us to expect some relation between EN and GMRES. In the following section, this relation is identified more explicitly.

A quantitative comparison of the experiments shows that $\|r_k\|_2$ in EN is larger than $\|r_{2k}\|_2$ in GMRES. Furthermore, in Figure 1 we observe, for $\alpha = 0$ and $\beta = 0.9$, peaks at $k = 9$ and $k = 50$, whereas in Figure 2 these peaks occur at $k = 16$ and $k = 79$. For the other situations similar observations have been made. This indicates that if GMRES leads to a peak at the kth iterate and EN to a peak at the jth iterate, then j is larger than $k/2$. This

TABLE 1
NUMBER OF ITERATION STEPS FOR WHICH
$\|r_i\|_2 / \|b\|_2 \leq 10^{-12}$

	Steps	
γ	EN	GMRES
0	38	65
30	44	84
60	35	70
300	86	150
3000	408	455

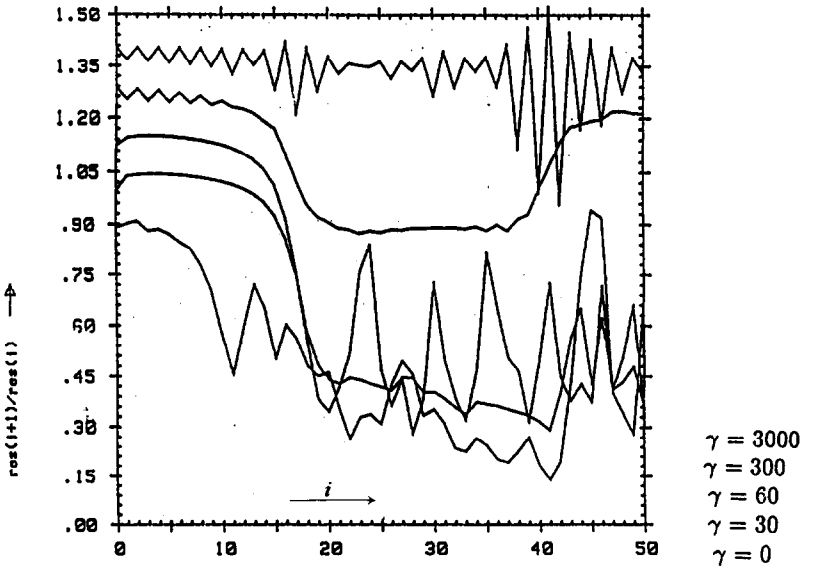

FIG. 9. Problem P10, EN.

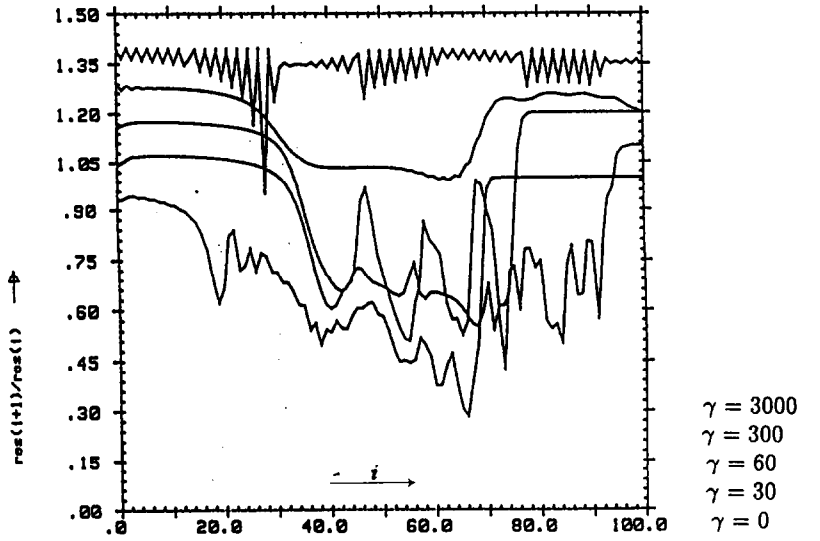

FIG. 10. Problem P10, GMRES.

again underlines the idea that the convergence behavior of GMRES after $2k$ steps is at least comparable with EN after k steps. This seems reasonable, since EN is more than twice as expensive per iteration step as GMRES. In these experiments the implementations EN1 and EN2 give the same results.

Finally we describe some numerical experiments for a more realistic problem. We take Ω to be the unit square and consider the pde

$$-\Delta u + \gamma u_x = 1 \quad \text{on } \Omega \quad \text{and} \quad u|_{\partial\Omega} = 0.$$

Using the standard five-point central finite-difference approximation over a uniform rectangular grid, we obtain a linear system (Problem P10). We take the step size in the x- and y-directions equal to $\frac{1}{30}$ [EN is applied to the system multiplied by $450/(\gamma/60+1)$]. Starting with $x_0 = (0,\ldots,0)^T$ gives the results shown in Table 1. Except for the choice $\gamma = 3000$, it appears from Table 1 that roughly $2k$ steps of GMRES are comparable with k steps of EN (see also Figures 9 and 10).

3. A COMPARISON OF EN AND GMRES

In this section we will show that the space spanned by the vectors c_k, generated by EN, is contained in a Krylov subspace. Furthermore, we will compare the norms of the residuals in EN and GMRES. Then by estimating the required amount of work and memory we will be able to compare the efficiency of the two methods.

First we will show that the vectors c_k which are generated by the EN method are elements of a Krylov subspace.

THEOREM 3.1. *If H_k is not singular and $E_k r_k \neq 0$, then*

$$r_k = r_0 + \sum_{i=1}^{2k} \alpha_{ki} (AH_0)^i r_0$$

and

$$\operatorname{span}\{c_0,\ldots,c_k\} \subset \operatorname{span}\{(AH_0)r_0,\ldots,(AH_0)^{2k+2}r_0\}.$$

Proof. In order to simplify relations, we redefine c_k, in this proof, as

$$c_k = Au_k \tag{3.1}$$

(note that only the direction of c_k is relevant).

We prove the theorem by an induction argument in k. From (3.1) it follows that $c_0 = AH_0E_0r_0 = AH_0(I - AH_0)r_0$, so that $c_0 \in \text{span}\{(AH_0)r_0, (AH_0)^2r_0\}$. This implies that the theorem is true for $k = 0$.

Combination of (1.4) and (1.5) gives

$$r_{k+1} = E_{k+1}r_k = (I - P_k)(I - AH_0)r_k = (I - AH_0)r_k - P_kE_0r_k.$$

Since P_k is the orthogonal projection onto $\text{span}\{c_0, \ldots, c_k\}$, it follows by induction that

$$r_{k+1} = r_0 + \sum_{i=1}^{2(k+1)} \alpha_{k+1,i}(AH_0)^i r_0. \tag{3.2}$$

Furthermore, from (3.1) we obtain $c_{k+1} = AH_{k+1}E_{k+1}r_{k+1} = (I - E_{k+1})E_{k+1}r_{k+1}$. Together with (1.4) this gives

$$c_{k+1} = [I - (I - P_k)(I - AH_0)]E_{k+1}r_{k+1} = (AH_0 + P_kE_0)E_{k+1}r_{k+1}.$$

Another application of (1.4) leads to

$$c_{k+1} = P_kE_0E_{k+1}r_{k+1} + AH_0(I - P_k)(I - AH_0)r_{k+1}$$

and hence

$$c_{k+1} = P_kE_0E_{k+1}r_{k+1} - AH_0P_kE_0r_{k+1} + AH_0(I - AH_0)r_{k+1}.$$

Since P_k is the orthogonal projection onto $\text{span}\{c_0, \ldots, c_k\}$, it follows by induction and (3.2) that $c_{k+1} \in \text{span}\{(AH_0)r_0, \ldots, (AH_0)^{2(k+1)+2}r_0\}$, which completes the proof. ■

The following definition is used for the comparison of the residuals of EN and GMRES.

TABLE 2
AMOUNT OF WORK AND MEMORY FOR DIFFERENT METHODS

Method	Steps	Multiplications with H_0	Multiplications with A	Inner products	Vector updates	Memory
EN1	k	$2k$	$4k$	k^2	k^2	$2kn$
EN2	k	$2k$	$2k$	k^2	$2k^2$	$2kn$
GMRES	$2k$	$2k$	$2k$	$2k^2$	$2k^2$	$2kn(+2k^2)$

DEFINITION 3.2. r_k^{EN} is the residual in the kth step of EN. r_k^G is the residual in the kth step of GMRES applied to the postconditioned linear system $AH_0y = b$, where H_0 is the same matrix in both methods (note that $x = H_0y$ solves the system $Ax = b$).

From Theorem 3.1 and (1.2) we obtain the following inequality:

$$\|r_k^{EN}\|_2 \geq \|r_{2k}^G\|_2. \tag{3.3}$$

This inequality supports our earlier observation on the numerical experiments reported in Section 2.

In order to compare the efficiency of EN and GMRES we need an estimate for the amount of work and memory in each method. For obvious reasons we have listed in Table 2 the amount of work and memory requirements for k steps of EN and $2k$ steps of GMRES.

The inner products in EN1 can be computed in parallel. Furthermore in EN2 the vector updates used to form η and ξ (or u_k and c_k) can be computed in parallel. The inner products and vector updates in the implementation of GMRES as given in Van der Vorst (1989) cannot be computed in parallel. This might be a disadvantage for GMRES in a parallel computing environment.

Since in most of our numerical experiments $\|r_k^{EN}\|_2$ and $\|r_{2k}^G\|_2$ differ considerably, we also give estimates for the amount of work and memory requirements for the following experiment. The solution of Problem P10 with $\gamma = 300$ is computed with the EN method and the GMRES method. The results are plotted in Figure 11. Note that EN requires more multiplications with H_0 and A than GMRES to obtain the same accuracy. Choosing eps $= 10^{-12}$, it appears that $\|r_{86}^{EN}\|_2 / \|b\|_2 \leq$ eps and $\|r_{150}^G\|_2 / \|b\|_2 \leq$ eps. The amount of work and memory requirements to obtain this accuracy are listed in Table 3.

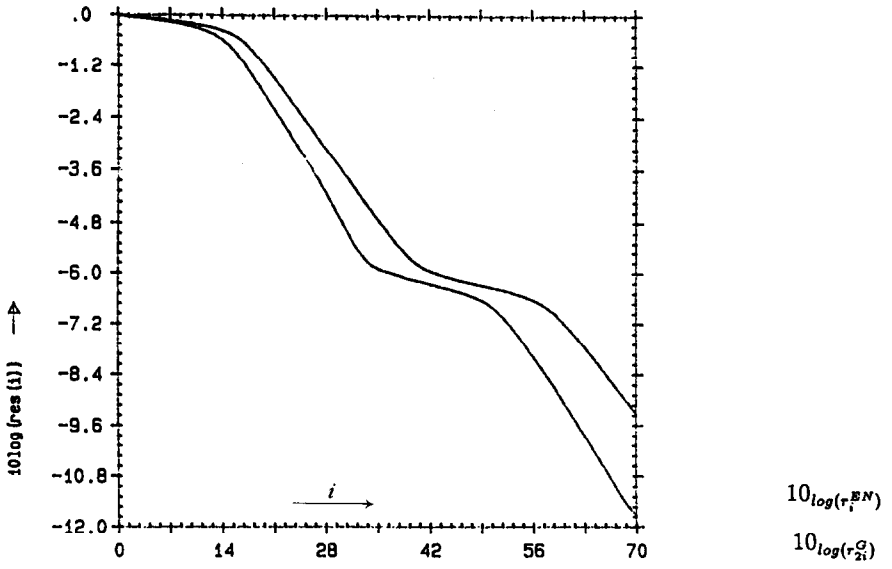

FIG. 11. Problem P10.

In practical situations the order of the linear system, n, will be much larger than the required number of iterations. In such cases the term $2k^2$ in the required amount of memory for the GMRES method is negligible.

We conclude that when $\|r_k^{EN}\|_2 \approx \|r_{2k}^{G}\|_2$, the EN2 method is more efficient than the GMRES method in terms of flops counts. However, the experiment has shown that there are problems for which $\|r_k^{EN}\|_2 \approx \|r_j^{G}\|_2$ with $j < 2k$. In the following section we will give more evidence for such situations. In such cases it is less clear which method is preferable in terms of flops counts. With respect to the memory requirements we note that GMRES is preferable.

TABLE 3

AMOUNT OF WORK AND MEMORY FOR DIFFERENT METHODS

Method	Steps	Multiplications with		Inner products	Vector updates	Memory
		H_0	A			
EN1	86	172	344	7396	7396	154800
EN2	86	172	172	7396	14792	154800
GMRES	150	150	150	11250	11250	135000(+11250)

4. SOME SPECIFIC PROPERTIES OF EN

In this section we will show that the convergence and stability properties of the EN method are not scaling invariant. Subsequently we will provide some examples where the EN method does not converge. Finally we will show that Property 1.6 is useless from a practical point of view.

4.1 *The Convergence Behavior of EN with Respect to Scaling*

From its construction it follows that GMRES is scaling invariant, which means that when the method is applied to the system $\rho Ax = \rho b$, then the iterates are the same for every choice of $\rho \neq 0$. One might expect from the foregoing that EN has the same property. However, from our experiments it follows that EN is not scaling invariant. This is well illustrated by the results for Problem P6 (with $\alpha = 10^{-5}$ and $\beta = 0.9$). In our first experiment we take $H_0 = \rho I$ as an approximation for A^{-1}. Obvious choices for ρ are $\rho = 1/\lambda_1$, $\rho = 2/(\lambda_1 + \lambda_n)$, and $\rho = 1/\lambda_n$, where $\lambda_1 = 1$ is the smallest and $\lambda_n = 100$ the largest eigenvalue of A. We obtain $\|r_{100}^{\text{EN}}\|_2 / \|r_0\|_2 = 10^{65}$ for $\rho = 1/\lambda_1 = 1$, $\|r_{38}^{\text{EN}}\|_2 / \|r_0\|_2 \leqslant 10^{-12}$ for $\rho = 2/(\lambda_1 + \lambda_2) = \frac{2}{101}$, and $\|r_{40}^{\text{EN}}\|_2 / \|r_0\|_2 \leqslant 10^{-12}$ for $\rho = 1/\lambda_n = \frac{1}{100}$. So the convergence behavior of EN strongly depends on the choice of ρ.

As a second experiment we apply EN to Problem P6 with $B := \rho B$ for $\rho = 10^{-1}$, 10^{-2}, 10^{-3}, and 10^{-4} and $H_0 = I$. The method is terminated as soon as $\|r_i\|_2 / \|b\|_2 \leqslant 10^{-12}$. The number of iteration steps, for different choices of ρ, is given in Table 4.

The convergence behavior is displayed in Figure 12. In this figure, each curve is plotted at the right scale. For $\rho = 10^{-1}$ we notice that initially the residuals increase. For $\rho = 10^{-2}$ the curve is identical to the corresponding curve in Figure 1. Note that the curves for $\rho = 10^{-3}$ and $\rho = 10^{-4}$ are nearly the same. Furthermore, these curves show a striking resemblance to the corresponding curve for GMRES in Figure 2.

TABLE 4

NUMBER OF ITERATION STEPS FOR WHICH $\|r_i\|_2 / \|b\|_2 \leqslant 10^{-12}$ (FOR P6)

ρ	Iterates
10^{-1}	78
10^{-2}	40
10^{-3}	64
10^{-4}	66

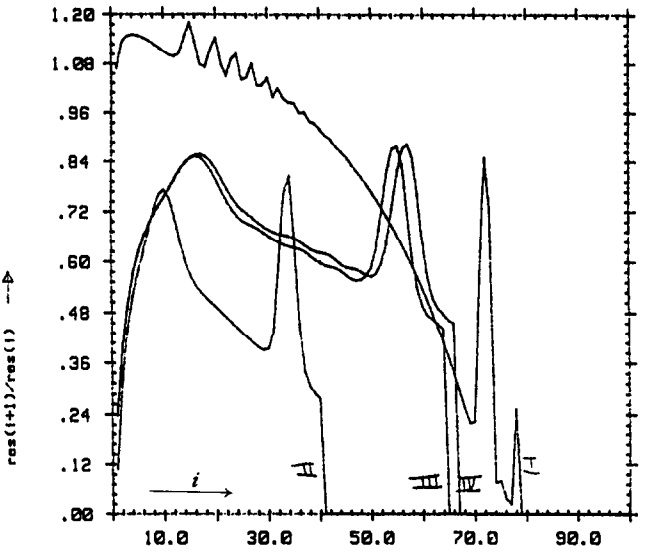

FIG. 12. Problem P6. I, $\rho = 10^{-1}$, II, $\rho = 10^{-2}$, III, $\rho = 10^{-3}$, IV, $\rho = 10^{-4}$.

An explanation of this might come from the observation that for $\rho = 10^{-4}$ we have that $E_0 = I - AH_0 \approx I$. This together with (1.4) implies

$$E_k \approx I - P_{k-1}.$$

Using this expression and (1.5), it follows from

$$u_k = H_k E_k r_k = H_k E_k^2 r_{k-1} \approx H_k E_k r_{k-1} = H_k r_k$$

that $u_k \approx H_k r_k$. This explains the resemblance of the curves, since the choice $u_k = H_k r_k$ leads to a method algebraically equivalent to GMRES [see Eirola and Nevanlinna (1989, p. 513) and also the following section].

In our example the choice $\rho = 10^{-2}$ is obviously preferable. We will call this value ρ_{opt} for our experiment. However, in general we know of no criterion which could be used for defining *a priori* an optimal ρ. Hence ρ_{opt} has to be determined experimentally. Furthermore, for this example we observe that for $\rho = 10\rho_{\text{opt}}$ the speed of convergence is halved, whereas for $\rho = 0.1\rho_{\text{opt}}$ the speed of convergence is approximately the same as for GMRES. Taking into account the amount of work and memory for the two methods

(see Table 2, Section 3) we conclude that we need a fairly good guess for ρ_{opt} if we want EN be more efficient than GMRES.

From these experiments it seems likely that the spectral radius of $I - AH_0$ has to be less than one (compare Section 4.2). This conjecture is confirmed by the following experiment. We take Ω to be the unit square and consider the pde

$$\Delta u = 0 \quad \text{on } \Omega \quad \text{and} \quad u|_{\partial\Omega} \text{ is given.}$$

Using the standard five-point central finite-difference approximation over an equidistant rectangular grid, we obtain a symmetric linear system. For H_0 we take an average of the incomplete Choleski (IC) and a modified incomplete Choleski (MIC) matrix; see Van der Vorst (1990, Section 3). The IC matrix corresponds to $\theta = 0$, whereas the MIC matrix corresponds to $\theta = 1$. Taking 200 points in the x- and y-directions, and $x_0 = 0$, we obtain the results given in Table 5.

Note that EN converges rather fast for the choices $0 \leqslant \theta \leqslant 0.98$ but diverges for the choice $\theta = 1$ which corresponds to the MIC preconditioner. This seems to be quite in line with similar experiments reported for preconditioned cg in Van der Vorst (1990, Section 3). However, if we apply EN to $0.1AH_0$ and $\theta = 1$, then we obtain $\|r_{23}^{\text{EN}}\|_2 / \|r_0\|_2 \leqslant 10^{-6}$. Therefore we believe that these experiments confirm our conjecture, since the spectral radius of $I - AH_0$ with the IC matrix is less than one, whereas with the MIC matrix the spectral radius is much larger than one (see Gustafsson, 1978). This result suggest that the divergence for $\theta = 1$ in the previous experiment is not caused only by a loss of independence among the Krylov subspace basis vectors for this value of θ [which is the reason for the slow convergence

TABLE 5

NUMBER OF ITERATION STEPS FOR WHICH $\|r_i\|_2 / \|r_0\|_2 \leqslant 10^{-6}$

θ	Iterates
0	21
0.5	17
0.9	14
0.95	13
0.96	14
0.97	13
0.98	17
0.99	46
1	*

of cg in this case (Van der Vorst, 1990, Section 3)]. We conclude that the convergence behavior of EN depends not only on the choice of H_0 but also on the scaling parameter ρ_{opt}. We expect good convergence if the spectral radius of $I - \rho_{opt} A H_0$ is less than one.

Our experiments show that EN is not invariant with respect to a general transformation of coordinates. Note that this conclusion is not in contradiction with Eirola and Nevanlinna (1989, Proposition 2.2), which states that EN is invariant under unitary transformations.

4.2. *The Stability of EN with Respect to Scaling*

From Figure 12 it appears that initially the residuals increase for $\rho = 10^{-1}$. To illustrate this phenomenon we will describe some experiments for ρ in the vicinity of 0.1. The results are given in Table 6, where i is the smallest value such that $\|r_i\|_2 / \|b\|_2 \leqslant 10^{-12}$ and i_{max} is defined by $\|r_{i_{max}}\|_2 = \max_{1 \leqslant j \leqslant i} \|r_j\|_2$.

This table shows that initial residuals increase fast for $\rho \leqslant 0.1$ and that the inequality $\|b - Ax_i\|_2 / \|b\|_2 \leqslant 10^{-12}$ does not hold for $\rho \geqslant 0.10$, as it should in exact arithmetic. For a possible explanation of the increase of the residuals we make use of the equality $r_{k+1} = (I - P_k) E_0 r_k$. The right-hand side consists of two parts: firstly a multiplication with E_0 and secondly a multiplication with the orthogonal projection $I - P_k$. Since $\sigma(E_0) \subset [1 - 100\rho, 1 - \rho]$, it follows that when $\rho > 2 \times 10^{-2}$, $\|E_0 r_k\|_2$ can be larger than $\|r_k\|_2$. For the second part we always have $\|(I - P_k) E_0 r_k\|_2 \leqslant \|E_0 r_k\|_2$. From this it appears that for $\rho \in (0.0, 0.02)$ the residual decreases in both parts. For $\rho \in [0.02, 0.09]$ the increase in the first part is canceled by the decrease in the second part. For $\rho \in (0.09, \infty)$, initially the increase in the first part dominates, whereas after a number of iteration steps (i_{max}) the decrease in the second part dominates.

Note that in exact arithmetic $r_i = b - Ax_i$. However, for $\rho \geqslant 0.1$ this is clearly violated in EN, and hence the reliability of r_i given by EN depends on the value of ρ. To explain this we let r_i and x_i denote the exact values, and

TABLE 6

$\|r_{i_{max}}\|_2$ FOR DIFFERENT VALUES OF ρ

ρ	i_{max}	$\|r_{i_{max}}\|_2 / \|b\|_2$	i	$\|r_i\|_2 / \|b\|_2$	$\|b - Ax_i\|_2 / \|b\|_2$
0.09	1	1	74	9×10^{-13}	9×10^{-13}
0.10	32	1.9×10^{1}	78	2.6×10^{-13}	4.3×10^{-13}
0.11	42	1.8×10^{3}	80	4.3×10^{-13}	2.9×10^{-11}
0.13	53	4.5×10^{7}	84	3.3×10^{-13}	1.4×10^{-6}
0.15	59	9.4×10^{11}	91	4.4×10^{-13}	2.2×10^{-2}

$\hat{r}_i$ and $\hat{x}_i$ denote the numerically computed values. Now define $\hat{z}_i = \hat{r}_{i_{\max}} - \hat{r}_i$ and $z_i = r_{i_{\max}} - r_i$, and suppose that $\|\hat{r}_{i_{\max}} - r_{i_{\max}}\|_2 / \|r_{i_{\max}}\|_2 \approx \epsilon$ and $\|\hat{z}_i - z_i\|_2 / \|z_i\|_2 \approx \epsilon$, where ϵ is a modest multiple of the machine precision. For $\rho = 0.15$ this implies $\|\hat{r}_i - r_i\|_2 = \|\hat{r}_{i_{\max}} - r_{i_{\max}} - (\hat{z}_i - z_i)\|_2 \approx \epsilon\{\|r_{i_{\max}}\|_2 + \|z\|_2\} \approx 2 \times 10^{12}\|b\|_2\epsilon$ and $\|\hat{r}_i - (b - A\hat{x}_i)\|_2 = \|\hat{r}_i - r_i + (b - Ax_i) - (b - A\hat{x}_i)\| \approx 2 \times 10^{12}\|b\|_2\epsilon$. This implies that due to rounding errors it is possible, for $\rho = 0.15$, that $\|r_i\|_2 / \|b\|_2 \leqslant 10^{-12}$ whereas $\|b - A\hat{x}_i\|_2 / \|b\|_2 \approx 10^{12}\epsilon$ [note that $\kappa_2(A) = 100$].

We conclude that the stability of the EN method depends on ρ. In the given experiment the EN method is quite stable for $\rho \leqslant 0.09$ and rather unstable for $\rho \geqslant 0.1$. It is, in general, not known for which ρ EN is stable. These results do not support the stability properties claimed in Eirola and Nevanlinna (1989, p. 516).

4.3. *Some Examples Where the EN Method Does Not Converge*

In this subsection we give some examples for which EN fails to converge. In order to identify such problems we look for nonsingular matrices A and H_0 such that H_1 is singular. Taking $H_0 = I$, it follows from (1.3) that if $E_0 r_0 \neq 0$ then $c_0 = \gamma A E_0 r_0$ with $\gamma = 1 / \|AE_0 r_0\|_2$. Using Property 1.6, it follows that H_1 is singular if and only if $c_0^T E_0 r_0 = \gamma (AE_0 r_0)^T E_0 r_0 = 0$. Thus A should be such that $(Av)^T v = 0$ for $v \in \mathbb{R}^n$, which means that Av and v are orthogonal. A simple matrix with this property is

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

EXAMPLE 1. We apply EN to $Ax = b$ with

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \qquad H_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$x = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Starting with

$$x_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

gives

$$r_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Since

$$E_0 = I - AH_0 = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix},$$

we obtain

$$E_o r_0 = \begin{pmatrix} 2 \\ 0 \end{pmatrix} \quad \text{and} \quad c_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

which implies that $c_0^T E_0 r_0 = 0$ (H_1 singular). Continuing the method with $H_1 = H_0$ yields $E_1 = E_0$ and $c_1^T E_0 r_1 = 0$ (H_2 singular). After k iteration steps we obtain $H_k = H_0$, $E_k = E_0$, and $r_k = E_0^k r_0$. The eigenvalues of E_0 are $1+i$ and $1-i$, so that

$$r_k = P \begin{bmatrix} (1+i)^k & 0 \\ 0 & (1-i)^k \end{bmatrix} P^{-1} r_0 \text{ and } \|r_k\|_2 \to \infty \qquad \text{for} \quad k \to \infty.$$

Thus, for this example the EN method is clearly divergent.

This example shows that EN does not converge for each given linear system. It is known that GMRES converges slowly for this type of matrices. In Huang and Van der Vorst, (1989, p. 23) it is shown that when GMRES is applied to $Ax = b$ with $A \in \mathbb{R}^{n \times n}$ given by

$$A = \begin{bmatrix} 0 & \cdot & \cdot & \cdot & \cdot & 0 & 1 \\ 1 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ & & \cdot & \cdot & & \cdot & \\ & & & \cdot & \cdot & & \cdot \\ & & & & \cdot & \cdot & \cdot \\ 0 & & & & & 1 & 0 \end{bmatrix}, \qquad b = (1,0,\ldots,0)^T, \quad \text{and}$$

$$x_0 = (0,\ldots,0)^T,$$

then $x_i = x_0$, $0 \leqslant i \leqslant n-1$, and $x_n = x$.

In our next example EN converges badly, whereas GMRES converges very fast.

EXAMPLE 2. Take

$$H_0 = I, \qquad A = \rho \begin{bmatrix} 0 & -10^4 & & & & 0 \\ 10^4 & 0 & & & & \\ & & 1.03 & & & \\ & & & 1.04 & & \\ & & & & \ddots & \\ 0 & & & & & 2 \end{bmatrix}, \quad \text{and}$$

$$x = (1, \ldots, 1)^T$$

Starting with $x_0 = (0, \ldots, 0)^T$, we obtain $\|r_{100}^{\text{EN1}}\|_2 / \|b\|_2 \geqslant 10^{-7}$ for $\rho = 10^{-3}$, whereas $\|r_{15}^{\text{G}}\|_2 / \|b\|_2 \leqslant 10^{-12}$. The rather bizarre convergence behavior of EN in dependence on the scaling parameter ρ is nicely illustrated by the fact that $\|r_{14}^{\text{EN1}}\|_2 / \|b\|_2 \leqslant 10^{-12}$ for $\rho = 10^{-4}$ but $\|r_{100}^{\text{EN1}}\|_2 / \|b\|_2 \geqslant 10^{-5}$ for $\rho = 10^{-5}$. Using the EN2 implementation, we obtain for the updated residual $\|r_{14}^{\text{EN2}}\|_2 / \|b\|_2 \leqslant 10^{-12}$ for $\rho = 10^{-3}$, 10^{-4}, and 10^{-5}, whereas the exact residual $\|Ax_{14} - b\|_2 / \|b\|_2$ equals 3×10^{-4}, 2×10^{-6}, and 5×10^{-7} for ρ respectively 10^{-3}, 10^{-4}, and 10^{-5}.

In Section 4.1 we have seen, and explained, that for ρ small enough, application of EN to $\rho Ax = \rho b$ gives $\|r_i^{\text{EN}}\|_2 \approx \|r_i^{\text{G}}\|_2$ for some problems. Example 2 shows that there are also linear systems where this equivalence does not hold.

4.4. The Practical Relevance of Property 1.6

In this subsection we consider the application of EN1 to Example 2 for $\rho = 10^{-3}$. Taking into account the similarity between Examples 1 and 2, we expect that in Example 2, H_1 is nearly singular. By computation it follows that $\|E_1 r_1\|_2 = 1.4 \times 10^2$ and $\|H_1 E_1 r_1\|_2 = 2.6 \times 10^{-10}$, so $\|H_1^{-1}\|_2$ is very large. It appears that the computed vector $c_1 = \gamma A H_1 E_1 r_1$ has a large relative error and that $c_0^T c_1$ equals 1.5×10^{-4} instead of 0. This explains the bad convergence behavior of EN1 in Example 2. The large relative error in c_1 is also predicted by Eirola and Nevanlinna (1989, p. 515, Theorem 2.2) if we use that

$$c_H = \frac{\|E_1\|_2}{\|AH_1 E_1 r_1\|_2} \geqslant 3 \times 10^8.$$

This experience motivates us to investigate the practical applicability of Property 1.6. We note the following drawbacks:

(1) if $c_k^T E_0 r_k = 0$, then it is possible of course that the computed value of $c_k^T E_0 r_k \neq 0$;

(2) if $c_k^T E_0 r_k \neq 0$, then it is still possible that H_{k+1} is nearly singular.

To get around these difficulties we could replace the condition $c_k^T E_0 r_k = 0$ by

$$\frac{|c_k^T E_0 r_k|}{\|E_0 r_k\|_2} \leqslant \epsilon \qquad \text{for} \quad \epsilon \geqslant 0. \tag{4.1}$$

If the inequality (4.1) holds, we take $H_{k+1} = H_k$. However, this condition has certain disadvantages too. First of all, it is not clear which values of ϵ are feasible. Secondly, implementation of this condition does not help much in Example 2. In that case we have $|c_0^T E_0 r_0| / \|E_0 r_0\|_2 = 1.8 \times 10^{-12}$. If we take $\epsilon \leqslant 1.8 \times 10^{-12}$, then we obtain the same results as without this condition, whereas $\epsilon > 1.8 \times 10^{-12}$ leads to $H_k = H_0$ for $0 \leqslant k \leqslant 100$ and $\|r_{100}\|_2 / \|b\|_2 = 10^{100}$. Hence, for Example 2 there is no value of ϵ such that the ENI method combined with (4.1) is convergent.

This indicates that implementing Property 1.6 in this way is useless from a practical point of view.

From the given examples it follows that EN is not attractive if some of the matrices H_k are (nearly) singular. Therefore, it is important to know *a priori* when the matrices H_k are (nearly) singular. In Eirola and Nevanlinna (1989, Theorem 2.3) the following "safe" case is stated: if AH_0 is positive (or negative) definite, then $\|(AH_k)^{-1}\|_2 \leqslant 1/\mu$, where

$$\mu = \inf_{\|x\|_2 = 1} \frac{\left|(AH_0 x)^T x\right|}{\left(\|x\|_2^2 + \|AH_0 x\|_2^2\right)^{1/2}}.$$

The following theorem states that if AH_0 is neither positive nor negative definite, then it is possible to obtain a singular matrix H_k.

THEOREM 4.2. *If AH_0 is neither positive nor negative definite on* $\mathrm{Im}(E_0)$, *then there exists a right-hand-side vector b such that H_1 is singular.*

Proof. The condition on AH_0 implies that there is a vector $v \in \mathrm{Im}(E_0)$ such that $(AH_0 v)^T v = 0$. Since $v \in \mathrm{Im}(E_0)$, we can find $b \in \mathbb{R}^n$ such that

$E_0 b = v$. Applying EN to this system with $x_0 = (0,\ldots,0)^T$ yields $c_0 = \gamma AH_0 E_0 b$ with $\gamma = 1/\|AH_0 E_0 b\|_2$. From Property 1.6 and the equations

$$c_0^T E_0 r_0 = \gamma (AH_0 E_0 b)^T E_0 b = \gamma (AH_0 v)^T v = 0$$

it follows that H_1 is singular. ■

Note that if there is a vector v such that $(Av)^T v = 0$, then $1/\mu$ is infinite.

Our conclusion is that it is only "safe" to apply the EN method if AH_0 is positive or negative definite.

4.5. *A Scaling Invariant Version of the* EN *Method*

In Section 4.1 and 4.2 we have shown that the convergence and stability properties of EN are not scaling invariant. As a consequence of this one should estimate a parameter ρ_{opt} such that the spectral radius of $I - \mu_{\text{opt}} AH_0$ is less than one. In this section we modify the EN method so that parameter estimation is not longer required.

In Section 1 we have shown that $r_{k+1} = (I - AH_k) r_k - \mu_k A u_k$. Combination with $u_k = H_k (I - AH_k) r_k$ gives $r_{k+1} = (I - \mu_k AH_k)(I - AH_k) r_k$. Since $E_k = (I - AH_k) = (I - P_{k-1})(I - AH_0)$, r_{k+1} can also be written as $r_{k+1} = (I - \mu_k AH_k)(I - P_{k-1})(I - AH_0) r_k$.

Note that it is the multiplication by $I - AH_0$ which makes EN not scaling invariant. Using this observation, we modify EN so that r_{k+1} obtained with the modified EN method can be written as follows:

$$r_{k+1} = (I - \mu_k AH_k)(I - P_{k-1})(I - \gamma_k AH_0) r_k,$$

where the constant $\gamma_k = (AH_0 r_k)^T r_k / \|AH_0 r_k\|_2^2$ minimizes $\|(I - \gamma AH_0) r_k\|_2$. To implement the modified method (EN3) the first step of the implementation EN2 should be changed as follows:

1. $\gamma = (AH_0 r_k)^T r_k / \|AH_0 r_k\|_2^2$, $\xi^{(0)} = (I - \gamma AH_0) r_k$, $\eta^{(0)} = \gamma H_0 r_k$, $\alpha_i = c_i^T \xi^{(i)}$, $\xi^{(i+1)} = \xi^{(i)} - \alpha_i c_i$, $\eta^{(i+1)} = \eta^{(i)} + \alpha_i u_i$, $i = 0,\ldots,k-1$.

It is easy to show that EN3 is scaling invariant, and that is confirmed by our numerical experiments.

Application of EN3 to Problem P6 (with $\alpha = 10^{-5}$ and $\beta = 0.9$) with $B := \rho B$ gives $\|r_{35}^{\text{EN}}\|_2 / \|r_0\|_2 \leqslant 10^{-12}$ for all choices of ρ. Finally we apply EN3 to the pde problem given in Section 4.1. The results are given in Table 7. Note that EN3 converges also for the choice $\theta = 1$. Furthermore, the optimal number of iterates of EN2 in Table 5 equals 13, whereas the optimal number of iterates of EN3 in Table 7 equals 9. Thus in this example we

TABLE 7
NUMBER OF ITERATION STEPS FOR WHICH $\|r_i^{EN3}\|_2 / \|r_0\|_2 \leq 10^{-6}$

θ	Iterates
0	21
0.5	18
0.9	12
0.95	11
0.96	11
0.97	10
0.98	10
0.99	9
1	22

observe that the convergence of EN3 is approximately 1.5 times as good as the convergence of EN2.

5. ANOTHER FORMULATION OF THE GMRES METHOD

In Eirola and Nevanlinna (1989, p. 513) it is noted without proof that, when choosing

$$u_k = H_k r_k, \tag{5.1}$$

instead of $u_k = H_k E_k r_k$, the EN method leads to an algorithm algebraically equivalent to GMRES. In this section we first prove this equivalence under the assumption that the matrices H_k are nonsingular. Subsequently we give a slight modification of the choice (5.1) such that the method remains equivalent to GMRES even if the matrices H_k are singular. A suitable implementation of this method arises if an orthonormal basis for the Krylov subspace is generated by the modified Gram-Schmidt process.

First we will show that the vectors c_k form an orthonormal basis for $\text{span}\{(AH_0)r_0, \ldots, (AH_0)^{k+1} r_0\}$. Since (1.4) is only valid for the choice $u_k = H_k E_k r_k$, we use the equality

$$E_{k+1} = \left(I - c_k c_k^T\right) E_k \tag{5.2}$$

[cf. Eirola and Nevanlinna (1989, p. 512)].

THEOREM 5.1. *Let u_k be chosen as $u_k = H_k r_k$ in* EN. *When H_k is not singular and $r_k \neq 0$, then $r_{k+1} = (I - P_k)r_0$, where $P_k = \sum_{i=0}^{k} c_i c_i^T$ is the orthogonal projection onto span$\{(AH_0)r_0, \ldots, (AH_0)^{k+1}r_0\}$.*

Proof. As in the proof for Theorem 3.1, we take

$$c_k = Au_k \tag{5.3}$$

We prove the theorem by an induction argument on k. Using (5.1) and (5.3), we obtain $c_0 = (AH_0)r_0$. Combination of (1.5) and (5.2) gives $r_1 - E_1 r_0 = (I - c_0 c_0^T)E_0 r_0$. Since $c_0 = (AH_0)r_0$, it follows that $r_1 = (I - c_0 c_0^T)r_0$. This implies that the theorem is true for $k = 0$.

It follows from (5.1) and (5.3) that $c_{k+1} = AH_{k+1}r_{k+1} = (I - E_{k+1})r_{k+1}$. Equation (5.2) implies $E_{k+1} = (I - P_k)E_0$ and $c_{k+1} = [I - (I - P_k)(I - AH_0)](I - P_k)r_0$ by induction. The last equation can also be written as

$$c_{k+1} = (I - P_k)AH_0(I - P_k)r_0 = (I - P_k)\left(c_0 - \sum_{i=0}^{k} (AH_0)c_i c_i^T r_0\right).$$

By induction it follows that $(AH_0)c_i \in \text{span}\{c_0, \ldots, c_k\}$ for $i = 0, \ldots, k-1$; hence

$$c_{k+1} = -(I - P_k)AH_0 c_k c_k^T r_0. \tag{5.4}$$

Since c_k has a nonzero component in the direction of $(AH_0)^{k+1}r_0$, H_{k+1} is nonsingular, and $r_{k+1} \neq 0$, it follows that c_{k+1} has a nonzero component in the direction of $(AH_0)^{k+2}r_0$. Using (5.4), it follows by induction that $c_i^T c_{k+1} = 0$, $i = 0, \ldots, k$. Thus $\{c_0, \ldots, c_{k+1}\}$ is an orthonormal basis for span$\{(AH_0)r_0, \ldots, (AH_0)^{k+2}r_0\}$. Combining (1.5), (5.2), and (5.3), we obtain

$$r_{k+2} = E_{k+2}r_{k+1} = (I - c_{k+1}c_{k+1}^T)(I - AH_{k+1})r_{k+1},$$

so that

$$r_{k+2} = (I - c_{k+1}c_{k+1}^T)(r_{k+1} - c_{k+1}) = (I - c_{k+1}c_{k+1}^T)r_{k+1}.$$

By induction it follows that $r_{k+2} = (I - P_k - c_{k+1}c_{k+1}^T)r_0$, which concludes our proof. ∎

From this theorem we conclude that the method converges if the matrices H_k are nonsingular. Since GMRES leads to the solution within a finite number of iteration steps, we look for a modification of (5.1) such that the condition on H_k can be dropped. To this end we note that it follows from (5.4) that c_{k+1} is a unit vector in the direction of $(I - P_k)(AH_0)c_k$. Choose the vector u_k as follows:

$$\begin{aligned} u_0 &= H_0 r_0, \\ u_k &= u_{k-1} - H_k c_{k-1}, \qquad k \geqslant 1, \end{aligned} \tag{5.5a}$$

which implies

$$\begin{aligned} c_0 &= AH_0 r_0, \\ c_k &= -(I - P_{k-1}) AH_0 c_{k-1}. \end{aligned} \tag{5.5b}$$

It can be proven that $r_{k+1} = (I - P_k)r_0$, where $P_k = \sum_{i=0}^{k} c_i c_i^T$ is the orthogonal projection onto $\text{span}\{(AH_0)r_0, \ldots, (AH_0)^{k+1} r_0\}$. Furthermore, it is easy to show if $c_k \neq 0$ and $c_{k+1} = 0$ then $r_{k+1} = 0$. From this remark and (1.2) it follows that EN with u_k as in (5.5) is equivalent to GMRES applied to the postconditioned linear system $AH_0 y = b$.

An implementation of this method is:

1. $u_0 = H_0 r_0 / \|AH_0 r_0\|_2$, $c_0 = Au_0$, $k = 0$, $x_1 = x_0 + u_0 c_0^T r_0$, and $r_1 = r_0 - c_0 c_0^T r_0$;
2. while $\|r_{k+1}\|_2 > \text{eps}$ do $k := k + 1$, $c_k^{(0)} = AH_0 c_{k-1}$, $u_k^{(0)} = H_0 c_{k-1}$, $\alpha_i = c_i^T c_k^{(i)}$, $c_k^{(i+1)} = c_k^{(i)} - \alpha_i c_i$, $u_k^{(i+1)} = u_k^{(i)} - \alpha_i u_i$, $i = 0, \ldots, k-1$, $c_k = c_k^{(k)} / \|c_k^{(k)}\|_2$, $u_k = u_k^{(k)} / \|c_k^{(k)}\|_2$,
3. $x_{k+1} = x_k + u_k c_k^T r_k$ and $r_{k+1} = r_k - c_k c_k^T r_k$.

Note that the vectors c_k are made mutually orthogonal by the modified Gram-Schmidt process. In this implementation of GMRES, $2k$ iteration steps involve $2k$ multiplications with A and H_0, $2k^2$ inner products, $4k^2$ vector updates, and $4kn$ memory space. Comparing this with GMRES in Table 2, it follows that this implementation requires $2k^2$ vector updates and $2kn$ memory space extra.

Using the choice (5.5b), the GMRES method is formulated in the same way as the EN method. This correspondence gives some theoretical insight: for instance, also for GMRES a matrix H_k can be formed which approximates the inverse of A. With respect to flops counts and memory requirements we prefer the implementation of GMRES given in Saad and Schultz (1986) and Van der Vorst (1989).

6. A COMPARISON OF THE EN AND THE B METHOD

In this section we compare the EN method with the B method described in Broyden (1969). The B method is mostly used to solve nonlinear systems, but it can also be used to solve a linear system. The description of the B method indicates certain similarities to the EN method. However, a further investigation reveals essential differences. The main difference is that $r_k = r_0 + \sum_{i=1}^{2k} \alpha_{ki}(AH_0)^i r_0$ for the EN method, whereas $r_k = r_0 + \sum_{i=1}^{k} \beta_{ki}(AH_0)^i r_0$ for the B method. We conclude this section with a comparison of the B and the GMRES method.

From the descriptions of the EN and B methods in Section 1, we note the following correspondence: in both methods rank-one updates are used to construct a matrix H_k which is an approximation to the inverse of A (compare Broyden, 1969 and 1970). In order to make a more detailed comparison we use the following vectors:

DEFINITION 6.1.

$$u_k = -(H_k A H_k - H_k) r_k, \qquad v_k = \frac{H_k^T H_k r_k}{r_k^T H_k^T H_k A H_k r_k}, \qquad c_k = A u_k.$$

Note that u_k, v_k, etc., are different for the different methods. Since only the residuals for the methods will be compared, we have chosen to identify them by a superscript as in r_k^B (for Broyden), where necessary.

From the description of the B method, Definition 6.1, and the equations $p_k = H_k r_k$ and $y_k = r_k - r_{k+1} = r_k - b + A(x_k + H_k r_k) = A H_k r_k$ we deduce that

$$H_{k+1} = H_k + u_k v_k^T. \tag{6.1}$$

THEOREM 6.2. *If* $r_k^T H_k^T H_k A H_k r_k \neq 0$ *then* $r_k = r_0 + \sum_{i=1}^{k} \alpha_{ki}(AH_0)^i r_0$, *where*

$$\operatorname{span}\{c_0, \ldots, c_k\} \subset \operatorname{span}\{(AH_0) r_0, \ldots, (AH_0)^{k+2} r_0\}.$$

Proof. We prove the theorem by an induction argument on k. From Definition 6.1 it follows that $c_0 = A u_0 = -(AH_0)^2 r_0 + (AH_0) r_0$; hence $\operatorname{span}\{c_0\} \subset \operatorname{span}\{(AH_0) r_0, (AH_0)^2 r_0\}$. This implies that the theorem is true for $k = 0$.

Since $x_{k+1} = x_k + H_k r_k$ we have $r_{k+1} = (I - AH_k)r_k$. This together with Definition 6.1 and (6.1) yields

$$r_{k+1} = r_k - A\left(H_0 + \sum_{i=0}^{k-1} u_i v_i^T\right) r_k = r_k - AH_0 r_k - \sum_{i=0}^{k-1} c_i v_i^T r_k.$$

Now, it follows by induction that

$$r_{k+1} = r_0 + \sum_{i=1}^{k+1} \alpha_{k+1,i} (AH_0)^i r_0. \tag{6.2}$$

By Definition 6.1 we have that $c_{k+1} = -(AH_{k+1})^2 r_{k+1} + (AH_{k+1}) r_{k+1}$. Since $H_{k+1} = H_0 + \sum_{i=0}^{k} u_i v_i^T$, the following equation holds:

$$c_{k+1} = -\left(AH_0 + \sum_{i=0}^{k} c_i v_i^T\right)^2 r_{k+1} + \left(AH_0 + \sum_{i=0}^{k} c_i v_i^T\right) r_{k+1}.$$

This implies that

$$c_{k+1} = -(AH_0)^2 r_{k+1} + (AH_0) r_{k+1} + \sum_{i=0}^{k} c_i v_i^T \left(-AH_0 - \sum_{i=0}^{k} c_i v_i^T + 1\right) r_{k+1} - \sum_{i=0}^{k} (AH_0) c_i v_i^T r_{k+1}.$$

Using (6.2), it follows by induction that $c_{k+1} \in \text{span}\{(AH_0 r_0, \ldots, (AH_0)^{k+3} r_0\}$. ■

Theorem 6.2 together with (1.2) yields the following inequality: $\|r_k^{\text{B}}\|_2 \geqslant \|r_k^{\text{G}}\|_2$. From the numerical experiments given in Section 2 it follows that $\|r_k^{\text{G}}\|_2$ can be much larger than $\|r_k^{\text{EN}}\|_2$. Hence, the B and EN methods cannot be equivalent. In Broyden (1970) a generalization of the B method is given. In this method, the BG method, the update of H_k is as follows:

$$H_{k+1} = H_k - (H_k y_k - p_k) q_k^T / q_k^T y_k,$$

where $p_k = H_k r_k$, $y_k = AH_k r_k$, and q_k is an arbitrary vector subject only to the restriction that $q_k^T y_k \neq 0$. Note that with the choice $q_k = H_k^T p_k$ the BG method is equal to the B method. In the same way as for the B method it

follows that $\|r_k^{BG}\|_2 \geqslant \|r_k^G\|_2$; hence there is no choice of q_k such that BG and EN are equivalent.

It can be shown that for $q_k = E_k^T A u_k$, BG is algebraically equivalent to GMRES, which starts with $x_0 + H_0(b - Ax_0)$. Furthermore, it appears that BG with $q_k = E_k^T A u_k$ is a secant method (Broyden, 1970). However, the specific method given in Broyden (1970, p. 373) is different from GMRES.

In order to estimate the efficiency of the BG method we make a comparison with the GMRES method. With respect to the amount of work and memory for an implementation of the BG method we note that the kth step costs at least one multiplication with H_0 and A together with k inner products and $2k$ vector updates, whereas $2k$ vectors of length n must be stored in memory. This, in combination with the inequality $\|r_k^{BG}\|_2 \geqslant \|r_k^G\|_2$, yields that for every choice of q_k the BG method is less efficient than GMRES applied to the postconditioned system $AH_0 y = b$.

7. CONCLUSIONS

In this paper we have compared the methods GMRES, EN, and B. From this comparison it appears that in some numerical experiments EN takes less work than GMRES. However, a theoretical investigation shows that the efficiency of EN can be at most only slightly better than that of GMRES. Furthermore, the numerical experiments show that the convergence and stability of EN are not scaling invariant. However, we specify a new version of the EN method which is scaling invariant. The convergence behavior of this version seems to be better than that of the original EN method. Subsequently we gave a formulation of GMRES in the same spirit as the EN method. This correspondence gives theoretical insight, but in practical situations we prefer the implementation of the GMRES method as given in Saad and Schultz (1986) and Van der Vorst (1989).

Since the class of BG methods proposed in Broyden (1970) seems to be related to EN, this class is included in our comparison. We show that the EN method (with $u_k = H_k E_k r_k$) is not equivalent to any BG method. With respect to GMRES, a BG method is specified which is algebraically equivalent to GMRES.

REFERENCES

Broyden, G. C. 1969. A new method of solving nonlinear simultaneous equations, *Comput. J.* 12:94–99.

Broyden, G. C. 1970. The convergence of single-rank quasi-Newton methods, *Math. Comp.* 24:365–382.

Eirola, T. and Nevanlinna, O. 1989. Accelerating with rank-one updates, *Linear Algebra Appl.* 121:511–520.

Gustafsson, I. 1978. A class of first order factorization methods, *BIT* 18:142–156.

Huang, Y. and Van der Vorst, H. A. 1989. Some observations on the Convergence Behavior of GMRES. Report 89-09, Delft Univ. of Technology.

Saad, Y. and Schultz, M. H. 1986. GMRES: A generalized minimal residual algorithm for solving non symmetric linear systems, *SIAM J. Sci. Statist. Comput.* 7:856–869.

Van der Vorst, H. A. (1989). The convergence behavior of some iterative solution methods, in *Proceedings of the Fifth International Symposium on Numerical Methods in Engineering*, (Gruber, R., Periaux, J., and Shaw, R. P., Eds.), Springer-Verlag, Berlin, Vol. 1, pp. 61–72.

Van der Vorst, H. A. (1990). The convergence behavior of preconditioned CG and CG-S in the presence of rounding errors, in *Proceedings of the PCG Conference*, Nijmegen, 15–17 June 1989 (O. Axelsson and L. Yu. Kolotilina (Eds.), Lecture Notes in Mathematics, 1457, Springer-Verlag, Berlin, pp. 126–136.

Received 23 July 1990; final manuscript accepted 24 October 1990