SIAM J. SCI. COMPUT.
Vol. 15, No. 3, pp. 511–527, May 1994

ERROR-MINIMIZING KRYLOV SUBSPACE METHODS*

RÜDIGER WEISS†

Abstract. Iterative methods for the solution of linear systems are usually controlled by the observation of the norm of the residual. In reality, the error should be controlled, but the error is not available. The residuals and the errors are connected by the condition number of the system matrix. If the system is well conditioned, the decrease of the errors is closely connected to the decrease of the residuals. For these cases, Krylov subspace methods that minimize the residuals in the Euclidean norm or in the energy norm are powerful solution techniques. If the system is ill conditioned, the residuals can decrease while the errors increase. For these systems, arising even from very simple and commonly used differential equations, iterative methods that minimize the residuals may require a large number of iterations to reduce the errors. The user may be misled to stop the iteration too early by small residuals. Two families of error-minimizing Krylov subspace methods are proposed to overcome these difficulties. Each of them is suited for different problem types.

Principles for the design of generalized cg methods that minimize the error are derived from the geometric convergence behavior of generalized cg methods. These methods use the transposed system matrix multiplied by the system matrix as the iteration matrix. By this technique a fast convergence is achieved for matrices with clustered singular values and scattered eigenvalues.

A class of Krylov subspace methods minimizing the error by using the simple transposed matrix as the iteration matrix is proposed. Various realization possibilities are inherent in these generalized minimum error methods. The methods are analyzed theoretically. Common and related properties with generalized conjugate gradient methods are presented. These techniques should be preferred if the eigenvalues are more clustered than the singular values. The first promising tests for one distinct method are presented.

Key words. conjugate gradients, convergence, linear systems, error-minimizing methods, Krylov methods

AMS subject classifications. 65F10, 65F50, 40A05

1. Background.

The purpose of this paper is to present error-minimizing Krylov subspace methods for the solution of the linear system

$$Ax = b. \tag{1}$$

The matrix A is a real, square matrix of dimension n, i.e., $A \in \mathbb{R}^{n\times n}$, and $x, b \in \mathbb{R}^n$. In general, the matrix A is nonsymmetric and nonpositive definite. Let us assume A to be nonsingular.

We use the following notation for norms: Let Z be a symmetric, positive definite matrix; then the norm $||y||_Z$ of any vector $y \in \mathbb{R}^n$ is defined by $||y||_Z = \sqrt{y^T Z y}$. If Z is nonsymmetric and nonpositive definite, then $||y||_Z^2$ is mnemonic abbreviation for $y^T Z y$. $||y||_I$ is the Euclidean norm $||y||$.

Let $K_k(B, y) = \text{span}(y, By, \ldots, B^k y)$ be the Krylov space spanned by the matrix $B \in \mathbb{R}^{n\times n}$ and the vector $y \in \mathbb{R}^n$.

For any iterative method for the solution of (1) the residuals $r_k = Ax_k - b$ and the errors $e_k = x_k - x$ are connected by

$$r_k = Ae_k. \tag{2}$$

As a result of $||r_k|| = ||Ae_k|| \le ||A|| \cdot ||e_k||$ and $||e_k|| = ||A^{-1}r_k|| \le ||A^{-1}|| \cdot ||r_k||$, the following inequalities are valid:

$$\frac{1}{\kappa}\frac{||r_k||}{||r_0||} \le \frac{||e_k||}{||e_0||} \le \kappa\frac{||r_k||}{||r_0||} \le \kappa^2\frac{||e_k||}{||e_0||}, \tag{3}$$

*Received by the editors September 24, 1992; accepted for publication (in revised form) March 26, 1993.

†Numerikforschung für Supercomputer, Rechenzentrum der Universität Karlsruhe, Postfach 6980, 76128 Karlsruhe, Germany (weiss@rz.uni-karlsruhe.de).

where $\kappa = ||A|| \cdot ||A^{-1}||$ is the condition number of A. If κ is near to one, the norm of the relative residuals is strongly connected to the norm of the relative errors. In other words, if the residuals decrease sufficiently the errors will be reduced as well. For a large condition number the residuals can decrease while the norm of the errors either remains the same or even increases.

The next example shows that a separation of the norm of the errors from the norm of the residuals happens even for very simple and commonly used systems.

Example 1. Solve the following linear system resulting from the discretization of a one-dimensional Laplace equation:

$$\begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} \cdot x = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

The matrix has the eigenvalues $\lambda_k = 2 \cdot (1 - \cos \frac{k\pi}{n+1})$, where n is the dimension of the system and $1 \leq k \leq n$. Because the matrix is symmetric, the condition number is

$$\kappa = \frac{|\lambda_{\max}|}{|\lambda_{\min}|} = \frac{1 + \cos \frac{\pi}{n+1}}{1 - \cos \frac{\pi}{n+1}}.$$

The dimension of the system is 1000, resulting in a large condition number of approximately $4 \cdot 10^5$. The system was solved by the classical conjugate gradient (cg) method [8] and the GMRES method [11]. The starting guess is $x_0 = 0$. Figure 1 shows that the residuals decrease while the errors do not decrease for either method until 500 matrix-vector multiplications have been performed.

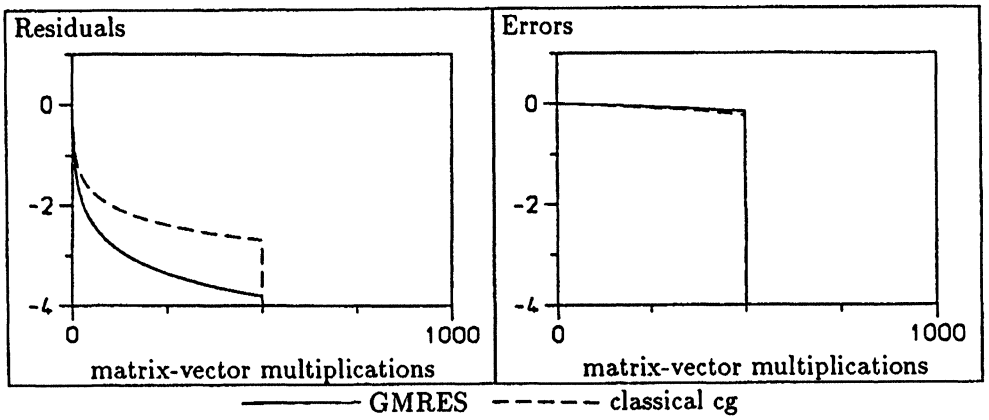

FIG. 1. *Norm of the relative residuals and errors of* GMRES *and classical cg (logarithmic scale).*

In the following sections we propose two remedies in order to obtain error-minimizing methods. The first technique is based on the iteration matrix $A^T A$ and generalized cg methods. The second technique uses the iteration matrix A^T and is a generalized Krylov subspace method.

2. Generalized conjugate gradient methods.

DEFINITION 1. *Let x_0 be any initial guess for the solution of the system $Ax = b$, $r_0 = Ax_0 - b$ the starting residual. The following recurrence is called a generalized cg method. Choose a preconditioning matrix P and calculate for $k \geq 1$ the residuals r_k and approximations x_k so that*

$$x_k \in x_0 + K_{k-1}(PA, Pr_0), \text{ with} \tag{4}$$

$$r_k^T Z r_{k-i} = 0 \tag{5}$$

for $i = 1, \ldots, \sigma_k$, where Z is an auxiliary, nonsingular matrix.

The method is called exact *if $\sigma_k = k$,* restarted *if $\sigma_k = (k-1) \bmod \sigma_{\text{res}} + 1$ with σ_{res} fixed,* truncated *if $\sigma_k = \min(k, \sigma_{\max})$ with $\sigma_{\max}$ fixed, and* combined *if the truncated method is restarted.*

From Definition 1 it follows directly for the residuals r_k and the errors $e_k = x_k - x$ that

$$r_k = \tilde{\Pi}_k(AP)r_0 = \sum_{i=1}^{k} \nu_{i,k}(AP)^i r_0 + r_0, \tag{6}$$

$$e_k = \tilde{\Pi}_k(PA)e_0 = \sum_{i=1}^{k} \nu_{i,k}(PA)^i e_0 + e_0, \tag{7}$$

where $\tilde{\Pi}_k$ is a polynomial of degree k with constant coefficient 1, i.e., $\tilde{\Pi}_k(0) = 1$.

If $Z = \tilde{Z}AP$ and $\tilde{Z}$ is symmetric, positive definite, then for exact and restarted methods (5) is equivalent to

$$\|r_k\|_{\tilde{Z}} = \min_{\mu_{1,k},\ldots,\mu_{\sigma_k,k}} \left\| \sum_{i=1}^{\sigma_k} \mu_{i,k}(AP)^i r_{k-\sigma_k} + r_{k-\sigma_k} \right\|_{\tilde{Z}}. \tag{8}$$

In this case the methods are called conjugate residual type methods or minimum residual methods.

If Z is symmetric, positive definite, then for exact and restarted methods (5) is equivalent to

$$\|\bar{r}_k\|_Z = \min_{\alpha_{1,k},\ldots,\alpha_{\sigma_k,k}} \left\| (AP)^{\sigma_k} r_{k-\sigma_k} + \sum_{i=1}^{\sigma_k} \alpha_{i,k}(AP)^{i-1} r_{k-\sigma_k} \right\|_Z, \tag{9}$$

where

$$\bar{r}_k = \frac{1}{\nu_{k,k}} r_k \tag{10}$$

is the pseudoresidual; see [12] and [14]. In this case the methods are called pseudoresidual methods.

The methods break down if division by zero occurs for the calculation of r_k. A breakdown may be curable, by modifying the algorithm [1], or incurable. We assume in the following that the algorithm does not break down.

LEMMA 2. *For any exact, generalized cg method*

$$r_k^T Z P^{-1} A^{-1} r_k = r_k^T Z P^{-1} A^{-1} \Pi_k(AP) r_0 \tag{11}$$

is satisfied for all matrix polynomials $\Pi_k(AP) = \sum_{i=1}^{k} \theta_i (AP)^i + I$ *(i.e.,* $\theta_1, \ldots, \theta_k$ *are arbitrary). This is especially true of*

$$r_k^T Z P^{-1} A^{-1} r_k = r_k^T Z P^{-1} A^{-1} r_j , \tag{12}$$

$$e_k^T A^T Z P^{-1} e_k = e_k^T A^T Z P^{-1} e_j \tag{13}$$

for $j = 0, \ldots, k$.

Proof. See [14]. □

The next investigations show some interesting facts concerning the geometric location of the residuals and the errors.

THEOREM 3. *The residuals* r_k *and the errors* e_k *of exact, generalized cg methods satisfy the following equations*:

$$\left\| r_k - \frac{\tilde{r}_j}{2} \right\|^2_{ZP^{-1}A^{-1}} = \frac{\|\tilde{r}_j\|^2_{ZP^{-1}A^{-1}}}{4}, \tag{14}$$

and

$$\left\| e_k - \frac{\tilde{e}_j}{2} \right\|^2_{A^T Z P^1} = \frac{\|\tilde{e}_j\|^2_{A^T Z P^{-1}}}{4} \tag{15}$$

for $j = 0, \ldots, k$ *with*

$$\tilde{r}_j = 2 \left(Z P^{-1} A^{-1} + (Z P^{-1} A^{-1})^T \right)^{-1} Z P^{-1} A^{-1} r_j, \tag{16}$$

$$\tilde{e}_j = 2 \left(Z P^{-1} + (Z P^{-1} A^{-1})^T A \right)^{-1} Z P^{-1} e_j. \tag{17}$$

In particular if $A^T Z P^{-1}$ *is symmetric, then*

$$\tilde{r}_j = r_j, \tag{18}$$

$$\tilde{e}_j = e_j. \tag{19}$$

Proof. By (12),

$$r_k^T Z P^{-1} A^{-1} (r_k - r_j) = 0.$$

Therefore,

$$r_k^T Z P^{-1} A^{-1} r_k - r_k^T Z P^{-1} A^{-1} r_j = 0.$$

From the definition (16) for $\tilde{r}_j$ follows

$$r_k^T Z P^{-1} A^{-1} r_k - \frac{1}{2} r_k^T Z P^{-1} A^{-1} A P Z^{-1} \left(Z P^{-1} A^{-1} + (Z P^{-1} A^{-1})^T \right) \tilde{r}_j = 0$$

and

$$r_k^T Z P^{-1} A^{-1} r_k - \frac{1}{2} r_k^T \left(Z P^{-1} A^{-1} + (Z P^{-1} A^{-1})^T \right) \tilde{r}_j + \frac{\|\tilde{r}_j\|^2_{ZP^{-1}A^{-1}}}{4} = \frac{\|\tilde{r}_j\|^2_{ZP^{-1}A^{-1}}}{4},$$

which is equivalent to

$$\left\| r_k - \frac{\tilde{r}_j}{2} \right\|^2_{ZP^{-1}A^{-1}} = \frac{\|\tilde{r}_j\|^2_{ZP^{-1}A^{-1}}}{4}.$$

Equation (15) follows by $r_j = Ae_j$. □

Equations (14) and (15) are quadratic forms and describe geometrical figures.

If $A^T ZP^{-1}$ is definite, then the residual r_k and the error e_k lie on hyperellipsoids. The lengths of the semiaxes of the hyperellipsoid for the residual are precisely the singular values of the matrix APZ^{-1} multiplied by $\|\tilde{r}_j\|_{ZP^{-1}A^{-1}}$. The lengths of the semiaxes of the hyperellipsoid for the error are precisely the singular values of the matrix $PZ^{-1}A^{-T}$ multiplied by $\|\tilde{e}_j\|_{A^T ZP^{-1}}$; see Fig. 2 for $j = 0$.

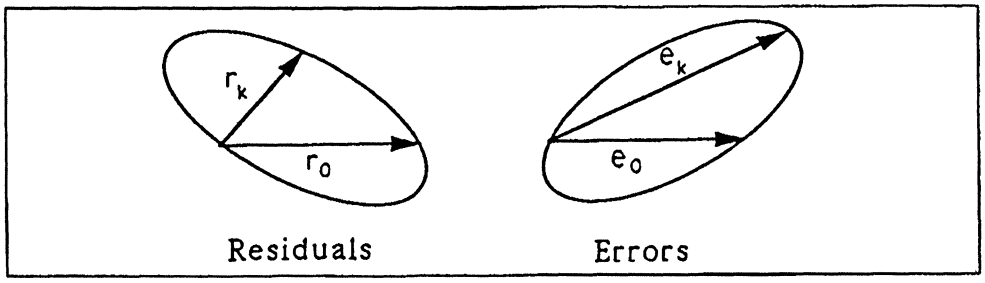

FIG. 2. *Residuals and errors if $A^T ZP^{-1}$ is definite.*

If $Z = AP$, then the residual r_k lies on an n-dimensional sphere; see Fig. 3 for $j = 0$. The norm of the residual is monotonically decreasing.

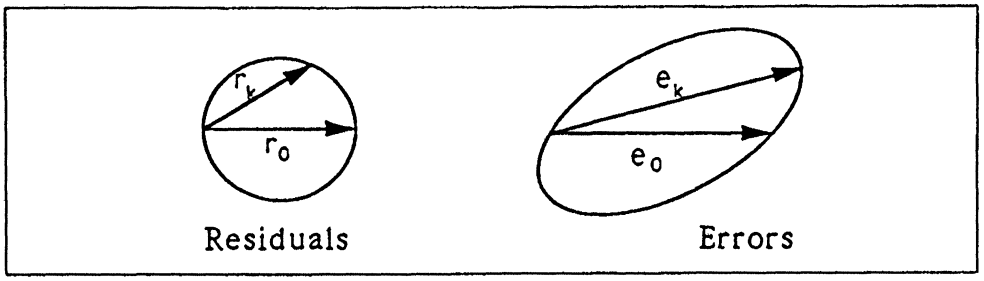

FIG. 3. *Residuals and errors if $Z = AP$.*

If $P = A^T Z$, then the error e_k lies on an n-dimensional sphere; see Fig. 4 for $j = 0$. The norm of the error is monotonically decreasing.

If $A^T ZP^{-1}$ is indefinite, then the geometric figures are not closed and in general the method does not converge; see Fig. 5 for $j = 0$.

The qualitative convergence behavior follows from Theorem 3 and the geometric interpretation. The speed of convergence depends on the eigenvalue distribution of the matrices AP and PA, respectively, which follows from the next lemma. Recall that the symmetric part of a matrix B is $\frac{1}{2}(B + B^T)$ and the skew-symmetric part is $\frac{1}{2}(B - B^T)$.

LEMMA 4. *If APZ^{-1} is positive real, i.e., the symmetric part of APZ^{-1} is positive definite,*

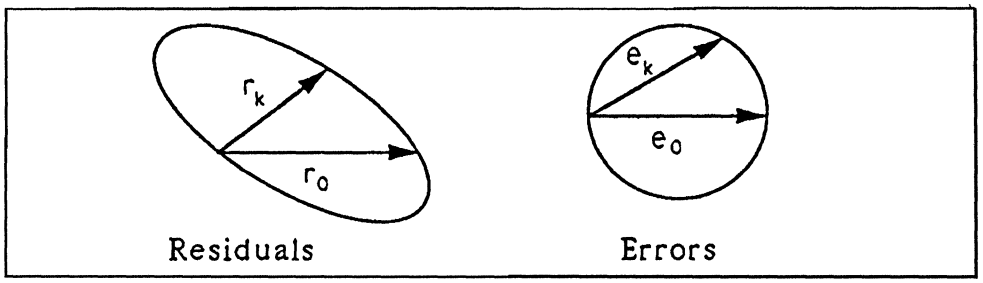

FIG. 4. *Residuals and errors if* $P = A^T Z$.

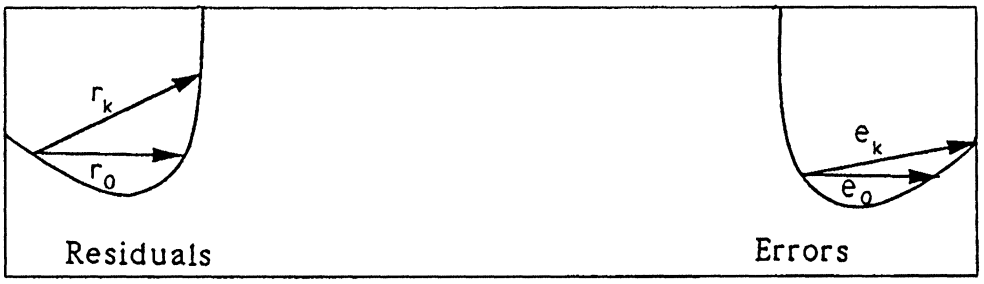

FIG. 5. *Residuals and errors if* $A^T Z P^{-1}$ *is indefinite.*

then

$$\|r_k\|_{ZP^{-1}A^{-1}} \leq \sqrt{1 + \frac{\rho^2(R)}{\mu_m^2}} \min_{\Pi_k} \|\Pi_k(AP)r_0\|_{ZP^{-1}A^{-1}}, \tag{20}$$

$$\|e_k\|_{A^T ZP^{-1}} \leq \sqrt{1 + \frac{\rho^2(R)}{\mu_m^2}} \min_{\Pi_k} \|\Pi_k(PA)e_0\|_{A^T ZP^{-1}} \tag{21}$$

holds for exact generalized cg methods. Π_k *is a polynomial of degree* k *with* $\Pi_k(0) = 1$. $\rho(R)$ *is the spectral radius of the skew-symmetric part* R *of* $ZP^{-1}A^{-1}$. μ_m *is the minimum eigenvalue of* M, *the symmetric part of* $ZP^{-1}A^{-1}$. *For residual-minimizing methods* ($Z = AP$)

$$\|r_k\| = \min_{\beta_1,\ldots,\beta_k} \left\| \sum_{i=1}^{k} \beta_i (AP)^i r_0 + r_0 \right\|. \tag{22}$$

For error-minimizing methods ($P = A^T Z$)

$$\begin{aligned} \|e_k\| &= \min_{\beta_1,\ldots,\beta_k} \left\| \sum_{i=1}^{k} \beta_i (PA)^i e_0 + e_0 \right\| \\ &= \min_{\beta_1,\ldots,\beta_k} \left\| \sum_{i=1}^{k} \beta_i (A^T ZA)^i e_0 + e_0 \right\| \end{aligned} \tag{23}$$

is valid.

Proof. See [14]. □

As

$$x^T APZ^{-1}x = x^T Z^{-T} P^T A^T x = y^T A^T ZP^{-1} y$$

with $y = PZ^{-1}x$ for all x, APZ^{-1} is positive real if and only if $A^T ZP^{-1}$ is positive real. Therefore Lemma 4, in fact, covers all converging cases following from Theorem 3 and the geometric interpretation. As a result of Theorem 3 and Lemma 4 the methods can be classified. In Table 1 the different conditions for the minimized items are collected.

TABLE 1
Various generalized cg methods.

Condition	Minimized Item	Method
$ZP^{-1} = I$ and		
A sym., pos. def.	$\|\|r_k\|\|_{A^{-1}}$	minimum energy norm
$ZP^{-1}A^{-1}$ sym., pos. def.	$\|\|r_k\|\|_{ZP^{-1}A^{-1}}$	minimum $ZP^{-1}A^{-1}$-residual
Z sym., pos. def.	$\|\|\tilde{r}_k\|\|_Z$	pseudoresidual
$Z = \tilde{Z}AP$ and		
$\tilde{Z}$ sym., pos. def.	$\|\|r_k\|\|_{\tilde{Z}}$	minimum $\tilde{Z}$-residual
$Z = AP$	$\|\|r_k\|\|$	minimum residual
$Z = A^{-T}\tilde{Z}P$ and		
$\tilde{Z}$ sym., pos. def.	$\|\|e_k\|\|_{\tilde{Z}}$	minimum $\tilde{Z}$-error
$P = A^T Z$	$\|\|e_k\|\|$	minimum error

For systems with clustered eigenvalues, i.e., a small condition number, the relative norm of the errors is closely connected to the relative norm of the residuals, and the speed of convergence is fast. For systems with a large condition number error-minimizing generalized cg methods ($P = A^T Z$) guarantee that the errors do not increase. These methods are techniques used to prevent the separation of the errors from the residuals.

The speed of convergence is dependent on the eigenvalue distribution of the iteration matrices AP for the residuals, PA for the errors, respectively. For error-minimizing generalized cg methods these matrices are $AA^T Z$ and $A^T ZA$. If the eigenvalues of $A^T ZA$ are clustered, then error-minimizing generalized cg methods guarantee a fast convergence. An investigation for the choice of Z in order to obtain a fast convergence on supercomputers is given in [15].

For Example 1 the eigenvalues of the matrix $A^T A$ are more scattered than the eigenvalues of A because the matrix is symmetric. Therefore an error-minimizing generalized cg method with $Z = I$ (Craig's method CGNE [2]) would prevent the separation of the residuals from the errors, but the convergence would be very slow. As regards the CPU time, this would become even worse because one iteration step needs two matrix-vector multiplications. For this example and many others the simple iteration matrix A or A^T would generate a faster convergence. In the next section we try to construct such methods that minimize the error.

3. Generalized minimum error methods. In this section we introduce our second technique: generalized minimum error methods that use $A^T P$ as the iteration matrix instead of using $A^T ZA$, as in the previous section. The technique is a generalization of a method proposed by Fridman [6] for symmetric, positive definite matrices.

The construction of the methods is based on the following lemma.

LEMMA 5. *Let* $y_i \in \mathbb{R}^n$, $q_i = A^T y_i$ *for* $i = k - \sigma_k, \ldots, k-1$ *and*

$$x_k = \sum_{i=k-\sigma_k}^{k-1} \gamma_{i,k} q_i + x_{k-\sigma_k}. \tag{24}$$

Then the error of $Ax_k - b$ is minimized in the Euclidean norm, i.e.,

$$||e_k|| = \min_{\gamma_{k-\sigma_k,k},\ldots,\gamma_{k-1,k}} \left\| \sum_{i=k-\sigma_k}^{k-1} \gamma_{i,k} q_i + x_{k-\sigma_k} - x \right\|, \tag{25}$$

by the solution $(\gamma_{k-\sigma_k,k}, \ldots, \gamma_{k-1,k})$ of the linear system

$$\sum_{i=k-\sigma_k}^{k-1} \gamma_{i,k} q_i^T q_j = b^T y_j - x_{k-\sigma_k}^T q_j, \tag{26}$$

$j = k - \sigma_k, \ldots, k-1$. The following orthogonalities are especially valid for $j = k - \sigma_k, \ldots, k-1$:

$$e_k^T q_j = 0. \tag{27}$$

Proof. The error $||e_k||$ is minimized if

$$\frac{1}{2} \frac{\partial}{\partial \gamma_{j,k}} ||e_k||^2 = e_k^T q_j = 0 \quad \text{for} \quad j = k - \sigma_k, \ldots, k-1,$$

with

$$\begin{aligned} e_k^T q_j &= (x_k - x)^T q_j \\ &= \sum_{i=k-\sigma_k}^{k-1} \gamma_{i,k} q_i^T q_j - x^T q_j + x_{k-\sigma_k}^T q_j \\ &= \sum_{i=k-\sigma_k}^{k-1} \gamma_{i,k} q_i^T q_j - b^T y_j + x_{k-\sigma_k}^T q_j = 0 \end{aligned}$$

following from (26). □

The q_i are update directions for the iterate x_k, and the y_i are auxiliary vectors needed for the computation of the coeffients $\gamma_{i,k}$. By means of Lemma 5 a whole family of error-minimizing Krylov subspace methods can be derived.

DEFINITION 6. *Let x_0 be any initial guess for the solution of the system $Ax = b$. Choose an auxiliary starting vector $y_0 \neq 0$, $q_0 = A^T y_0$. The following recurrence is called a generalized minimum error method* (GMERR). *Choose a preconditioning matrix P and calculate the following for $k \geq 1$:*

$$x_k = \sum_{i=k-\sigma_k}^{k-1} \gamma_{i,k} q_i + x_{k-\sigma_k}. \tag{28}$$

Choose $y_k \in K_k(PA^T, y_0)$ so that

$$q_k^T Z q_{k-i} = 0 \tag{29}$$

for $i = 1, \ldots, \bar{\sigma}_k$, where Z is an auxiliary, nonsingular matrix and

$$q_k = A^T y_k. \tag{30}$$

The $\gamma_{i,k}$ are determined from

$$||e_k|| = \min_{\gamma_{k-\sigma_k,k},\ldots,\gamma_{k-1,k}} \left\| \sum_{i=k-\sigma_k}^{k-1} \gamma_{i,k} q_i + e_{k-\sigma_k} \right\| . \tag{31}$$

The method is called exact *if* $\bar{\sigma}_k = k$, restarted *if* $\bar{\sigma}_k = (k-1) \bmod \sigma_{\text{res}} + 1$ *with* σ_{res} *fixed,* truncated *if* $\bar{\sigma}_k = \min(k, \sigma_{\max})$ *with* $\sigma_{\max}$ *fixed, and* combined *if the truncated method is restarted. The method is called* consistent *if* $\sigma_k = \bar{\sigma}_k$.

The coefficients $\gamma_{i,k}$ can be calculated by (26). If $ZA^T P = P^T AZ$, then the iterates q_k can be calculated by a simple three-term recurrence, and the methods are equivalent to the classical algorithm of Fridman [6]. The optimal choice of y_0 would be $y_0 = A^{-T} e_0 = A^{-T} A^{-1} r_0$ because then $q_0 = e_0$, and the solution is obtained in the first iteration step.

From Definition 6 follow directly

$$||e_k|| \leq ||e_{k-1}|| , \tag{32}$$

$$x_k \in x_0 + K_{k-1}(A^T P, q_0), \tag{33}$$

$$r_k \in r_0 + A\, K_{k-1}(A^T P, q_0), \tag{34}$$

$$q_k \in K_k(A^T P, q_0) . \tag{35}$$

From Definition 6 and condition (34) follow for the residuals and the errors:

LEMMA 7. *For* GMERRs *the following hold*:

$$e_k = \sum_{i=0}^{k-1} \nu_{i,k} (A^T P)^i q_0 + e_0, \tag{36}$$

$$r_k = \sum_{i=0}^{k-1} \nu_{i,k} A\, (A^T P)^i q_0 + r_0 . \tag{37}$$

Proof. The proof is trivial. □

From Lemma 7 the norm of the error can be estimated.

THEOREM 8. *For exact, consistent,* GMERRs *holds*

$$||e_k|| = \min_{\theta_0,\ldots,\theta_{k-1}} \left\| \sum_{i=0}^{k-1} \theta_i (A^T P)^i q_0 + e_0 \right\| . \tag{38}$$

Proof. Equation (38) follows directly from (31) and (36). □

From (38) it follows directly that for $P = ZA$ and $y_0 = Zr_0$ we also obtain the error-minimizing generalized cg methods of the previous section as GMERRs (see Lemma 4):

$$||e_k|| = \min_{\beta_1,\ldots,\beta_k} \left\| \sum_{i=1}^{k} \beta_i (A^T ZA)^i e_0 + e_0 \right\| . \tag{39}$$

Thus we have the following duality: The preconditioning matrix $P = ZA$ for GMERRs (see (39)) corresponds to the preconditioning matrix $P = A^T Z$ of generalized cg methods (see (23)).

The next theorem shows the interconnections of GMERRs to a one-dimensional minimization technique, the smoothing algorithm.

THEOREM 9. *If $Z = I$, then the calculation of x_k simplifies for exact, consistent* GMERRs *to*

$$x_k = x_{k-1} + \gamma_{k-1,k} q_{k-1}, \tag{40}$$

with

$$\gamma_{k-1,k} = \frac{b^T y_{k-1} - x_{k-1}^T q_{k-1}}{q_{k-1}^T q_{k-1}}. \tag{41}$$

Proof. The orthogonalities in Equation (29) simplify (26). Therefore, $\gamma_{k-i,j} = \gamma_{k-i,l}$ for all j and l, and the $\gamma_{k-i,k}$ can be calculated explicitly for $i = 1, \ldots, \sigma_k$ by

$$\begin{aligned}\gamma_{k-i,k} &= \frac{b^T y_{k-i} - x_0^T q_{k-i}}{q_{k-i}^T q_{k-i}} \\ &= \frac{b^T y_{k-i} - (\sum_{j=0}^{k-i-1} \gamma_{j,k-i} q_j + x_0)^T q_{k-i}}{q_{k-i}^T q_{k-i}}\end{aligned}$$

because of the orthogonalities

$$= \frac{b^T y_{k-i} - x_{k-i}^T q_{k-i}}{q_{k-i}^T q_{k-i}}.$$

Thus $x_k = x_0 + \sum_{i=0}^{k-1} \gamma_{i,k} q_i = x_{k-1} + \gamma_{k-1,k} q_{k-1}$. □

Equations (41) and (40) can be considered a smoothing algorithm to minimize the error. Schönauer introduced a smoothing algorithm to minimize the residual (see, e.g., [12]). In [14] it is shown that this algorithm transforms generalized cg methods that minimize the pseudoresidual to methods that minimize the residual. Gutknecht [7] gives the reverse of this smoothing algorithm, which is itself a smoothing algorithm. Zhou and Walker [16] show that a smoothing algorithm transforms the BCG method [3], [9] to the QMR method [5] and that it transforms the CGS method [13] to TFQMR [4].

If $\gamma_{k-1,k}$ is calculated according to (41), then the algorithm will be stable in the sense that x_k is depending only on values of the previous iteration step.

A GMERR can be implemented as shown in the following algorithm:

Algorithm I. *Let x_0 be any initial guess. Choose an auxiliary starting vector y_0 and calculate $q_0 = A^T y_0$. Choose δ and $\bar{\delta}$ accordingly. For $k \geq 1$ choose a preconditioning matrix P and calculate*:

$$\gamma_{k-1,k} = \frac{b^T y_{k-1} - x_{k-1}^T q_{k-1}}{\|q_{k-1}\|^2}, \tag{42}$$

$$\text{if } |\gamma_{k-1,k}| \leq \delta \text{ for } k > 1,$$

then restart

$$x_k = x_{k-1} + \gamma_{k-1,k} q_{k-1}, \tag{43}$$

$$\alpha_{i,k} = -\frac{q_{k-i}^T Z A^T P q_{k-1}}{\|q_{k-i}\|_Z^2}, \quad \text{for } i = 0, \ldots, \sigma_k \tag{44}$$

$$\bar{q}_k = A^T P q_{k-1} + \sum_{i=1}^{\sigma_k} \alpha_{i,k} q_{k-i}\,, \tag{45}$$

$$\text{if } \|\bar{q}_k\|_Z \leq \bar{\delta},$$

then restart

$$\phi_k = \frac{1}{\|\bar{q}_k\|_Z}, \tag{46}$$

$$q_k = \phi_k \bar{q}_k, \tag{47}$$

$$y_k = \phi_k \left(P q_{k-1} + \sum_{i=1}^{\sigma_k} \alpha_{i,k} y_{k-i} \right). \tag{48}$$

This implementation is an Arnoldi-like algorithm for q_k. If we calculate

$$\phi_k = \frac{1}{\sum_{i=1}^{\sigma_k} \alpha_{i,k}}$$

instead of (46), then Algorithm I resembles an ORTHORES algorithm. Without the restart function the algorithm would break down if $\bar{q}_k = 0$. An invariant subspace or the whole space would then be spanned, and the iteration would be restarted from the defect correction equation.

The algorithm is also restarted if the iterate x_k does not change sufficiently. By numerical engineering the value of δ has been optimized to $\delta = 3 \cdot 10^{-3} \|x_{k-1}\|$. Note that the iteration restarts from the defect correction equation so that the iterate x_k is decreasing, thus justifying the choice of δ.

4. Comparison of residual-minimizing/error-minimizing methods. We test the convergence behavior of the residual-minimizing method GMRES [11], the error-minimizing generalized cg method CGNE [2], and a GMERR method. The intention is to get a feeling for how GMERR methods behave in comparison with the corresponding GMRES methods and the error-minimizing cg methods. GMRES is an exact generalized cg method that minimizes the Euclidean norm of the residuals ($Z = A$, $P = I$). Correspondingly, we select the exact GMERR method with $Z = P = I$ according to Algorithm I with $\bar{\delta} = 10^{-8}$ (14 digits accuracy). For CGNE holds $P = A^T$, $Z = I$. For all methods we choose $x_0 = 0$, and for GMERR $y_0 = r_0$.

Example 1 (continuation). As predicted, CGNE behaves worse than do GMRES and GMERR because the eigenvalues of the iteration matrix are more scattered. Unfortunately, GMERR does not converge faster than GMRES, but for GMERR the norm of the residuals is connected to the norm of the errors. Thus the user is not misled by the residuals, see Fig. 6. The bad convergence is dependent on the bad starting vector $x_0 = 0$.

Examples MIT1–MIT8. The examples are taken from [10]. We omit the first example MIT1 because the system matrix is the unit matrix and the solution is obtained in the first iteration step. The examples "nail down the space of matrices at every corner" [10]. The dimension of the system is 40 for MIT2, MIT3, MIT4, MIT6, and MIT7. The dimension of the system is 400 for MIT5 and MIT8. The solution was prescribed by the random number generator, and the right-hand side was calculated accordingly.

MIT2: A is a random matrix of dimension 40. All methods are similarly bad (see Fig. 7).

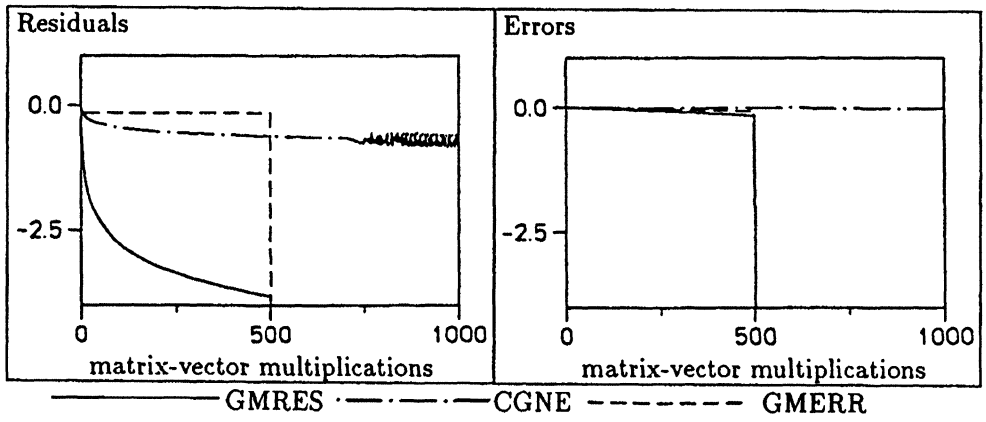

FIG. 6. *Norm of the relative residuals and errors of* GMRES, CGNE, *and* GMERR *for Example* 1 (*logarithmic scale*).

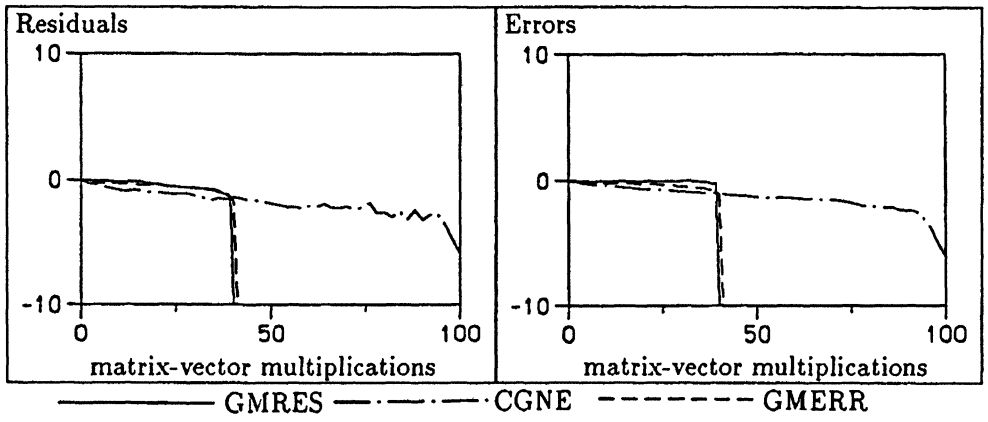

FIG. 7. *Norm of the relative residuals and errors of* GMRES, CGNE, *and* GMERR *for example* MIT2 (*logarithmic scale*).

MIT3: The matrix of MIT3 is

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & \ddots & 1 \\ 1 & 0 & \dots & \dots & 0 \end{pmatrix}.$$

GMERR beats GMRES and gets the solution immediately because $q_0 = A^T y_0 = A^T r_0 = A^T A e_0 = e_0$ has an optimal direction. CGNE converges immediately because $A^T A = I$ (see Fig. 8). The eigenvalues of A are scattered and the eigenvalues of $A^T A$ are clustered. Therefore, CGNE should be better than GMERR. The opposite is true because of the artificial and special choice of A.

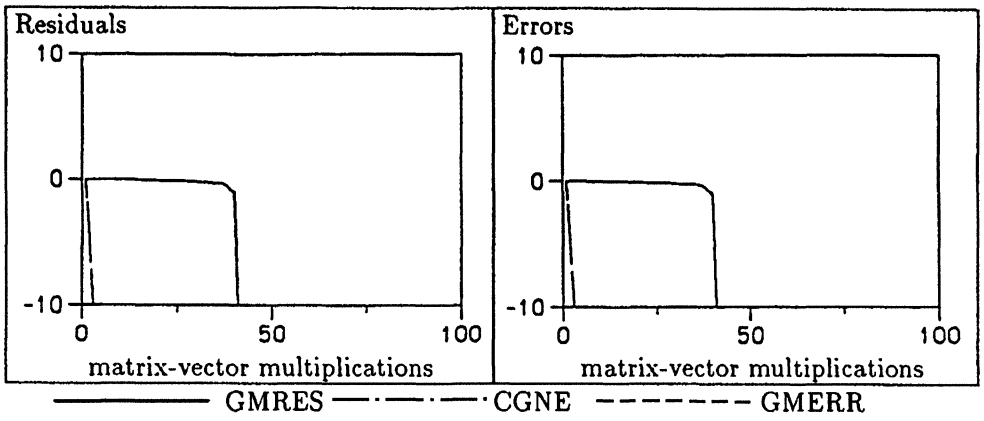

FIG. 8. *Norm of the relative residuals and errors of* GMRES, CGNE, *and* GMERR *for example* MIT3 (*logarithmic scale*).

The structure of the matrix for MIT4, MIT6, MIT7, and MIT8 is

$$A = \begin{pmatrix} A_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & A_{nn} \end{pmatrix}.$$

MIT4: $A_i = \begin{pmatrix} 1 & i-1 \\ 0 & 1 \end{pmatrix}$ The dimension is 40. GMRES is faster than GMERR and GMERR is faster than CGNE (see Fig. 9).

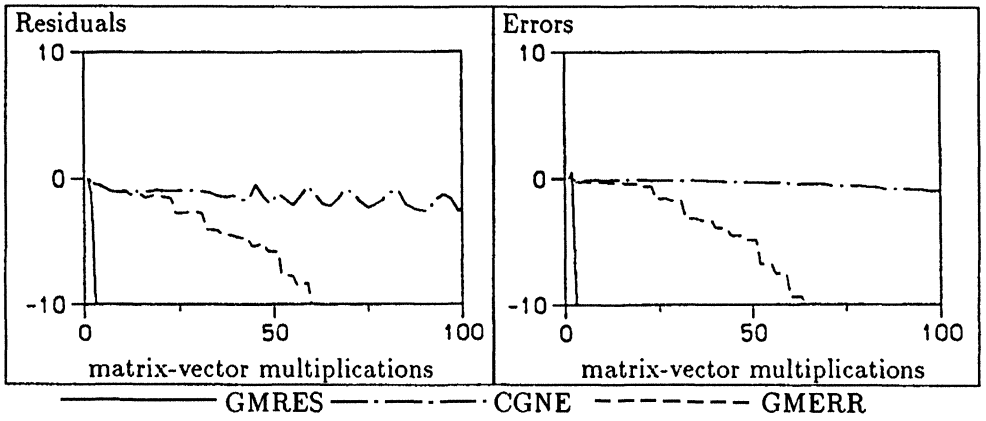

FIG. 9. *Norm of the relative residuals and errors of* GMRES, CGNE, *and* GMERR *for example* MIT4 (*logarithmic scale*).

MIT5:

$$A = \operatorname{diag}(\xi_1, \dots, \xi_n) \quad \text{with} \quad \kappa = \left(\frac{1 + \epsilon^{\frac{1}{2\sqrt{n}}}}{1 - \epsilon^{\frac{1}{2\sqrt{n}}}} \right)^2,$$

$\xi_i = 1 + \frac{1}{2}(y_i + 1)(\kappa - 1)$, $y_i = \cos\frac{(i-1)\pi}{n-1}$, and $\epsilon = 10^{-10}$. The dimension is 400. GMRES is faster than GMERR, and GMERR is faster than CGNE (see Fig. 10).

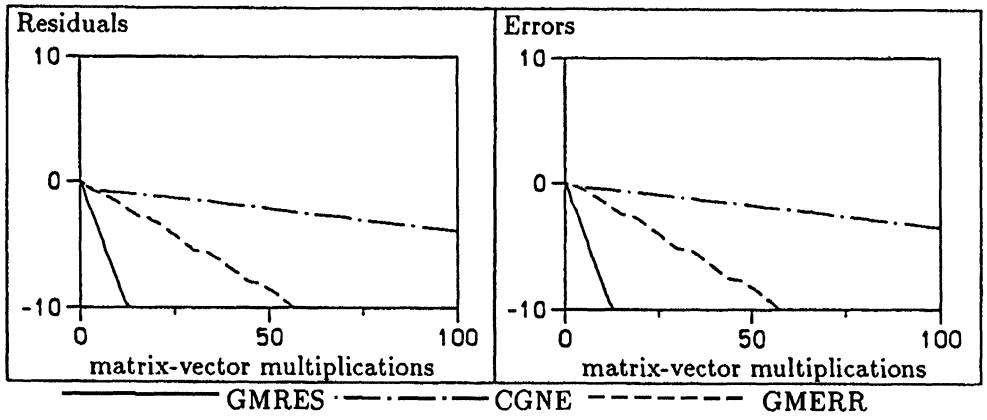

FIG. 10. *Norm of the relative residuals and errors of* GMRES, CGNE, *and* GMERR *for example* MIT5 *(logarithmic scale).*

MIT6: $A_i = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. The dimension is 40. All methods are very fast because $A^2 = A^{2T} = -I$ and $AA^T = I$ (see Fig. 11).

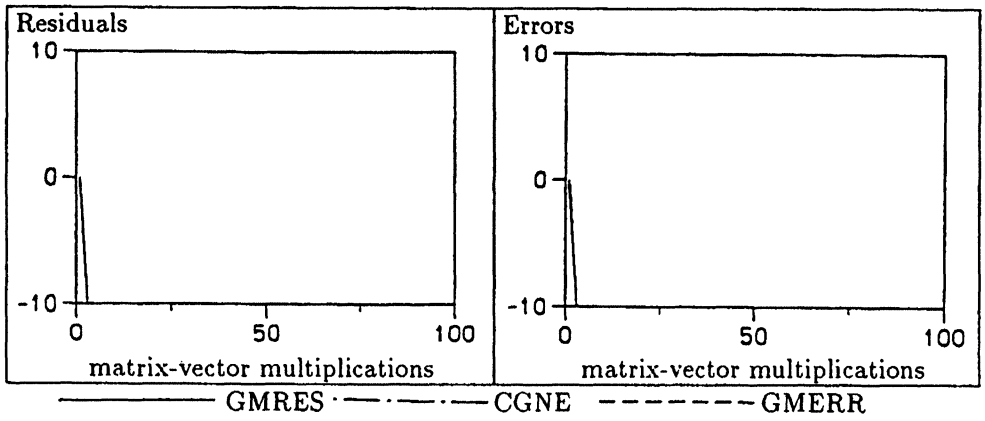

FIG. 11. *Norm of the relative residuals and errors of* GMRES, CGNE, *and* GMERR *for example* MIT6 *(logarithmic scale).*

MIT7: $A_i = \begin{pmatrix} 1 & i-1 \\ 0 & -1 \end{pmatrix}$. The dimension is 40. GMRES is very fast, GMERR is very slow, and CGNE is even worse (see Fig. 12).

MIT8:

$$A_i = \begin{pmatrix} \xi_i & \nu_i \\ 0 & \frac{\kappa}{\xi_i} \end{pmatrix} \quad \text{with} \quad \kappa = \left(\frac{1 + \epsilon^{\frac{\sqrt{nn}}{2}}}{1 - \epsilon^{\frac{\sqrt{nn}}{2}}} \right)^2,$$

$\xi_i = 1 + \frac{1}{2}(y_i + 1)(\kappa - 1)$, $y_i = \cos\frac{(i-1)\pi}{nn-1}$ and $\epsilon = 10^{-10}$. The dimension is 400. GMERR is twice as fast as GMRES; CGNE is as fast as GMERR (see Fig. 13).

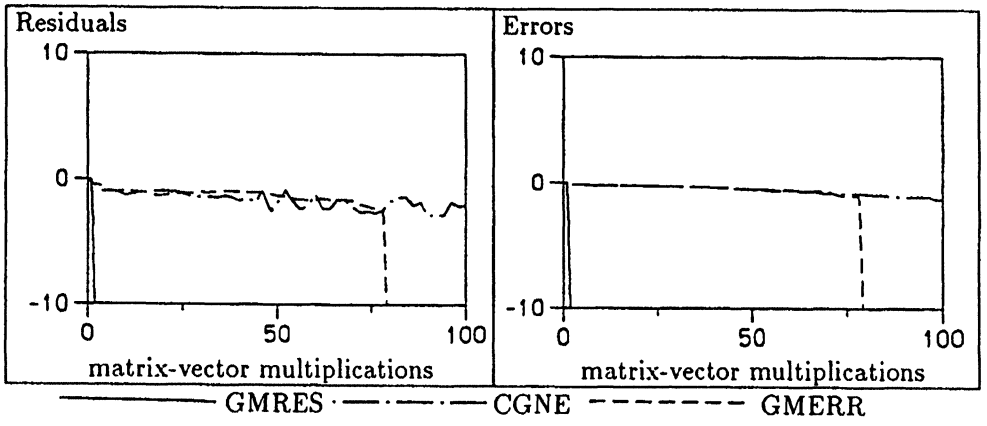

FIG. 12. *Norm of the relative residuals and errors of* GMRES, CGNE, *and* GMERR *for example* MIT7 *(logarithmic scale).*

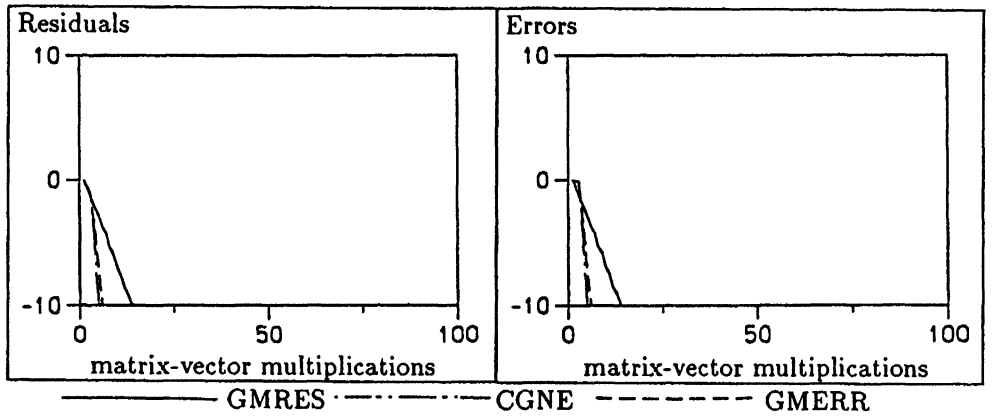

FIG. 13. *Norm of the relative residuals and errors of* GMRES, CGNE, *and* GMERR *for example* MIT8 *(logarithmic scale).*

For MIT2 and MIT6 GMRES and GMERR deliver comparable results. For MIT4, MIT5, and MIT7 GMRES is better; and for MIT3 and MIT8 GMERR beats GMRES. GMERR is better than CGNE for MIT4 and MIT5. For all other cases the two methods are comparable. In Table 2 the results are collected.

5. Outlook. The tests with the MIT examples indicate that exact, consistent GMERR according to Algorithm I is better than CGNE, and it is in some cases better than GMRES. But for more realistic problems, i.e., large problems arising from the discretization and linearization of partial differential equations, the exact methods are not feasible for storage requirements. Therefore, truncated and restarted versions have to be applied. While it was possible to optimize the restart and truncation parameters for GMRES-like methods, see e.g. [12], this optimization still lies ahead for GMERR methods. First tests indicate that this will be a more difficult task.

We further remark that GMERRs can be combined with generalized cg methods as we

TABLE 2
Comparison of GMRES, CGNE, *and* GMERR, *on a scale of* 1–3 (1=*best method*; 3=*worst method*).

Example	GMRES	CGNE	GMERR
MIT2	all methods bad		
MIT3	3	both 1	
MIT4	1	3	2
MIT5	1	3	2
MIT6	all methods good		
MIT7	1	both 3	
MIT8	3	both 1	

shall describe. For the biconjugate gradients (BCG) [3], [9] the double system $\hat{A}\hat{x} = \hat{b}$, i.e.,

$$\begin{pmatrix} A & 0 \\ 0 & A^T \end{pmatrix}\begin{pmatrix} x \\ x^* \end{pmatrix} = \begin{pmatrix} b \\ b^* \end{pmatrix},$$

is considered. b^* is arbitrary. The residuals have the form $\hat{r} = \binom{r}{r^*}$ and $Z = Z_B = \binom{0 \quad I}{I \quad 0}$. A generalized cg method is applied to this double system. As $Z_B\hat{A} = \hat{A}^T Z_B$, the method is exact for $\sigma = 2$, i.e., the sequence terminates automatically. Of course this doubling technique can be adapted to GMERRs in order to get terminating sequences. But we can also exploit this doubling differently. In the original BCG method the second sequence (*) is used only for the orthogonalization without exploitation of the iterates of the second sequence. But we can exploit the information of the second sequence for a generalized minimum error method especially because the iteration matrix of the second sequence is A^T as needed. Thus by modifying BCG slightly we obtain a generalized cg and a GMERR in parallel.

6. Conclusion. For problems with a single cluster of eigenvalues, i.e., in the symmetric case for well-conditioned problems, the errors decrease with the residuals, and the convergence of generalized cg methods is fast.

For problems with scattered eigenvalues, i.e., ill-conditioned problems in the symmetric case, the decrease of the residuals may be excellent while the errors do not decrease. As a consequence, we use error-minimizing methods. We have proposed two techniques. One is based on generalized cg methods with the iteration matrix $A^T ZA$, the other is based on GMERRs with the iteration matrix $A^T P$. We have shown various interconnections between the two techniques.

Because the speed of convergence is dependent on the eigenvalue distribution of the iteration matrix both techniques are suited for different problem types. The first is superior for scattered eigenvalues of the system matrix but more clustered eigenvalues of $A^T ZA$, i.e., for clustered singular values if $Z = I$. The second is preferable if the eigenvalues of $A^T ZA$, i.e., the singular values if $Z = I$, are more scattered than the eigenvalues of the system matrix.

REFERENCES

[1] C. BREZINSKI, M. R. ZAGLIA, AND H. SADOK, *Avoiding breakdown and near-breakdown in Lanczos type algorithms*, Numer. Algorithms, 1 (1991), pp. 261–284.

[2] E. J. CRAIG, *The n-step iteration procedures*, Math. Phys., 34 (1955), pp. 64–73.
[3] R. FLETCHER, *Conjugate gradient methods for indefinite systems*, in Proceedings of the Dundee Biennial Conference on Numerical Analysis, G. A. Watson, ed., Springer-Verlag, Berlin, New York, 1975, pp. 73–89.
[4] R. W. FREUND, *A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems*, SIAM J. Sci. Comput., 14 (1993), pp. 470–482.
[5] R. W. FREUND AND N. M. NACHTIGAL, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.
[6] V. M. FRIDMAN, *The method of minimum iterations with minimum errors for a system of linear algebraic equations with a symmetrical matrix*, U.S.S.R. Comput. Math. and Math. Phys., 2 (1963), pp. 362–363.
[7] M. H. GUTKNECHT, *Changing the norm in conjugate gradient type algorithms*, SIAM J. Numer. Anal., 30 (1993), pp. 40–56.
[8] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–435.
[9] C. LANCZOS, *Solution of systems of linear equations by minimized iterations*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 33–53.
[10] N. M. NACHTIGAL, S. C. REDDY, AND L. N. TREFETHEN, *How fast are nonsymmetric matrix iterations*? SIAM J. Matrix Anal. Appl., 13 (1992), pp. 778–795.
[11] Y. SAAD AND M. H. SCHULTZ, GMRES: *A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[12] W. SCHÖNAUER, *Scientific Computing on Vector Computers*, North-Holland, Amsterdam, New York, 1987.
[13] P. SONNEVELD, CGS, *a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 36–52.
[14] R. WEISS, *Convergence behavior of generalized conjugate gradient methods*, Internal Report 43/90, Computing Center, University of Karlsruhe, Karlsruhe, Germany, 1990.
[15] R. WEISS AND W. SCHÖNAUER, *Preconditioned conjugate gradient methods: What do we have and what do we need*? in Advances in Computer Methods for Partial Differential Equations, R. Vichnevetsky, D. Knight, and G. Richter, eds., International Association for Mathematics and Computers in Simulation, New Brunswick, NJ, 1992, pp. 806–812.
[16] L. ZHOU AND H. F. WALKER, *Residual smoothing techniques for iterative methods*, SIAM J. Sci. Comput., 15 (1992), pp. 297–312.