Journal of Computational Mathematics, Vol.16, No.6, 1998, 539–550.

TOTAL GENERALIZED MINIMUM BACKWARD ERROR ALGORITHM FOR SOLVING NONSYMMETRIC LINEAR SYSTEMS*1)

Zhi-hao Cao
(*Dearptment of Mathematics, Fudan University, Shanghai 200433, China*)

Abstract

This paper extendes the results by E.M. Kasenally[7] on a Generalized Minimum Backward Error Algorithm for nonsymmetric linear systems $Ax = b$ to the problem in which pertubations are simultaneously permitted on A and b. The approach adopted by Kasenally has been to view the approximate solution as the exact solution to a perturbed linear system in which changes are permitted to the matrix A only. The new method introduced in this paper is a Krylov subspace iterative method which minimizes the norm of the perturbations to both the observation vector b and the data matrix A and has better performance than the Kasenally's method and the restarted GMRES method[12]. The minimization problem amounts to computing the smallest singular value and the corresponding right singular vector of a low-order upper-Hessenberg matrix. Theoratical properties of the algorithm are discussed and practical implementation issues are considered. The numerical examples are also given.

Key words: Nonsymmetric linear systems, Iterative methods, Backward error.

1. Introduction

An important aspect of any iterative method for approximating the solution of a linear system

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

where A is an $n \times n$ real nonsymmetric matrix and b is an n-vector, is to decide at what point to stop the iteration. We customarily use the residual error as a stopping condition. The residual error $r_m = b - Ax_m$ can be viewed as a perturbation to the vector b such that the approximate solution is an exact solution of the perturbed linear system $Ax = b + \delta$, in which changes are permitted to the vector b only. The GMRES algorithm is based on classical Krylov subspace techniques and computes an approximate solution restricted to an affine space while minimising the backward perturbation norm of the vector b. From this backward error analysis of view E.M. Kasenally has viewed the approximate solution as an exact one of the perturbed linear system $(A - \Delta)x = b$,

* Received March 14, 1996.
1) Supported by the State Major Key Project for Basic Researches of China and the Doctral program of the China State Education Commission.

in which changes are permitted to the matrix A only. The Krylov subspace algorithm GMBACK proposed by Kasenally[7] computes an approximate solution restricted to an affine space while minimizing the backward perturbation norm of the matrix A. In this paper we view the approximate solution as an exact solution of the perturbed linear system $(A-\Delta)x = b+\delta$, in which changes are simultaneously permitted on matrix A and b[1,9,10]. A new Krylov subspace algorithm TGMBACK, which computes an approximate solution restricted to an affine space and minimizing the backward perturbation norm of the matrix A and vector b is presented. This minimization problem amounts to computing the smallest singular value and the corresponding right singular vector of a low-order upper Hessenberg matrix. The advantage for considering the algorithms which minimize the backward error is that there is often some uncertainty in the data A and b of the original linear systems and we can compare the backward error with the size of the uncertainty. Moreover, we found from numerical examples that the new method has better performance than Kasenally's method and restarted GMRES method.

The outline of this paper is as follows. Section 2 gives a backward error analysis for any iterative method for solving linear systems. The TGMBACK algorithm is introduced in Section 3. Some practical implementation issues and the numerical examples are presented in Section 4 and Section 5, respectively.

2. Backward Error Analysis for Iterative Methods

Consider the linear system in (1.1), where A is a large nonsymmetric matrix. Let $\{x_m\}$ be a sequence of approximate solutions produced by any iterative method. We first compare the residual error $r_m \equiv b - Ax_m$ with the minimum backward error $\Delta_{\min}$ in matrix A which satisfies $\|\Delta_{\min}\|_F = \min\{\|\Delta\|_F : (A-\Delta)x_m = b\}$.

Theorem 2.1. *Let x_m be an approximate solution of the linear system* (1.1) *and $\Delta_{\min}$ be the minimum backward error Δ in the matrix A such that $(A-\Delta)x_m = b$. Then*

$$\|\Delta_{\min}\|_F = \|r_m\|_2/\|x_m\|_2, \tag{2.1}$$

where $\|.\|_F$ is the Frobenious norm.

Proof. The residual equation

$$r_m = b - Ax_m$$

can be rewritten as follows

$$\Big(A + \frac{r_m x_m^T}{\|x_m\|_2^2}\Big)x_m = b$$

which implies that

$$\|\Delta_{\min}\|_F \le \|r_m x_m^T/\|x_m\|_2^2\|_F = \|r_m\|_2/\|x_m\|_2. \tag{2.2}$$

On the other hand, we have

$$(A-\Delta_{\min})x_m = b$$

which implies that

$$r_m = -\Delta_{\min} x_m,$$

therefore, we have

$$\|r_m\|_2 \le \|\Delta_{\min}\|_F \|x_m\|_2. \tag{2.3}$$

Combining (2.2) and (2.3) we deduce (2.1), thus completing the proof.

Remark 2.1. From the proof of Theorem 2.1 we have the following result:

$$\|\Delta_{\min}\|_2 = \|\Delta_{\min}\|_F = \|r_m\|_2/\|x_m\|_2. \tag{2.4}$$

Remark 2.2. From (2.1) and (2.4) we can see that if $\|r_m\|_2$ is small then the $\|\Delta_{\min}\|_F \equiv \|\Delta_{\min}\|_2$ is not necessarily small.

In order to derive the minimum perturbation both on matrix A and vector b, we need the notations of the Kroneker product $A \otimes B$ of matrice A and B, and the vec-function vec(A) of matrix A[8], and the following proposition (cf.[8] ch.12,sec.1)

Proposition 2.1. *If the orders of the matrices involved are such that all the operations bellow are defined, then*

$$(A \otimes B)(C \otimes D) = AC \otimes BD.$$
$$(A \otimes B)^T = A^T \otimes B^T.$$
$$vec(AXB) = (B^T \otimes A)vec(X).$$

We also need the following result (cf. [2] for proof).

Lemma 2.1. *Let $A \in R^{m\times n}$, $B \in R^{p\times q}$, $D \in R^{m\times q}$. Then the matrix equation*

$$AXB = D$$

is consistent if and only if for some $A^{(1)}$, $B^{(1)}$

$$AA^{(1)}DB^{(1)}B = D,$$

in which case the general solution of the matrix equation is

$$X = A^{(1)}DB^{(1)} + Y - A^{(1)}AYBB^{(1)}$$

for arbitrary $Y \in R^{n\times p}$. Here $A^{(1)}$ denotes the $\{1\}$-inverse of a matrix A.

Theorem 2.2. *Let x_m be an approximate solution of the linear system (1.1) and $[\Delta, \delta]_{\min}$ be the minimum backward error $[\Delta, \delta]$ in $[A, b]$ such that*

$$(A - \Delta)x_m = b + \delta. \tag{2.5}$$

Then

$$[\Delta, \delta]_{\min} = -r_m w_m^T/(1 + \|x_m\|_2^2)^{1/2}, \tag{2.6}$$

where $r_m = b - Ax_m$ and $w_m = [x_m^T, 1]^T/(1 + \|x_m\|_2^2)^{1/2}$. Therefore

$$\|[\Delta, \delta]_{\min}\|_F = \|[\Delta, \delta]_{\min}\|_2 = \|r_m\|_2/\sqrt{1 + \|x_m\|_2^2}. \tag{2.7}$$

Proof. We rewrite (2.5) as follows

$$\Delta x_m + \delta 1 = -r_m$$

which can be read as

$$\begin{pmatrix} \Delta & \delta \end{pmatrix} \begin{pmatrix} x_m \\ 1 \end{pmatrix} = -r_m. \tag{2.8}$$

We now use Lemma 2.1 to solve the matrix equation (2.8) for $[\Delta, \delta]$. Obviously, it holds

$$\begin{pmatrix} x_m \\ 1 \end{pmatrix}^{(1)} = \frac{[x_m^T, 1]}{1 + \|x_m\|_2^2} = \frac{w_m^T}{(1 + \|x_m\|_2^2)^{1/2}}.$$

Therefore

$$-r_m \begin{pmatrix} x_m \\ 1 \end{pmatrix}^{(1)} \begin{pmatrix} x_m \\ 1 \end{pmatrix} = -r_m \frac{[x_m^T, 1]}{1 + \|x_m\|_2^2} \begin{pmatrix} x_m \\ 1 \end{pmatrix} = -r_m.$$

Thus, from Lemma 2.1 the matrix equation is consistent and the general solution is

$$[\Delta, \delta] = -r_m \frac{w_m^T}{(1 + \|x_m\|_2^2)^{1/2}} + Y(I - w_m w_m^T). \tag{2.9}$$

Using Proposition 2.1 we deduce that

$$\|[\Delta, \delta]_{\min}\|_F^2 = \min_Y \Big\| [(I - w_m w_m^T) \otimes I] \mathrm{vec}(Y) - \mathrm{vec}\Big(r_m \frac{w_m^T}{(1 + \|x_m\|_2^2)^{1/2}} \Big) \Big\|_2^2. \tag{2.10}$$

The right-hand side of (2.10) is a least squares problem. Using proposition 2.1 again we can deduce its normal equation as follows

$$[(I - w_m w_m^T) \otimes I] \mathrm{vec}(Y) = [(I - w_m w_m^T) \otimes I] \mathrm{vec}\Big(r_m \frac{w_m^T}{(1 + \|x_m\|_2^2)^{1/2}} \Big), \tag{2.11}$$

which can be written back to the following form

$$Y(I - w_m w_m^T) - \frac{r_m}{1 + \|x_m\|_2^2} w_m^T (I - w_m w_m^T) = 0.$$

Therefore, we have

$$Y(I - w_m w_m^T) = 0. \tag{2.12}$$

Instituting (2.12) into (2.9) we deduce (2.6), thus completing the proof.

Remark 2.3. From Theorem 2.2 (cf.(2.7)) we can see that the norm of the total minimum perturbation $\|[\Delta, \delta]_{\min}\|_F = \|[\Delta, \delta]_{\min}\|_2$ is always smaller than the norm of the residual $\|r_m\|_2$. Thus, from the view of the backward error analysis[13] we can say that if both the matrix A and the vector b in the system include data errors, then using the size of the residual norm $\|r_m\|_2$ as a stopping criteria of an iterative method for solving the linear system (1.1) is reasonable.

Remark 2.4. If we use $\|D[\Delta, \delta]\|_F = \min$, where $D = \mathrm{diag}\,(d_1, \cdots, d_n)$ is a nonsingular diagonal matrix[5], then we have

$$(D[\Delta, \delta])_{\min} = -D r_m w_m^T / \|w_m\|_2^2$$

and

$$\|(D[\Delta,\delta])_{\min}\|_F = \|(D[\Delta,\delta])_{\min}\|_2 = \|Dr_m\|_2/\sqrt{1+\|x_m\|_2^2}.$$

3. The TGMBACK Algorithm

Let us briefly review the krylov subspace iterative methods for solving linear system (1.1). If x_0 is the initial solution estimate, then the initial residual is $r_0 = b - Ax_0$. Define the m-dimensional Krylov subspace $\mathcal{K}_m(A, r_0)$, select, for example by the Arnoldi process, the columns of the matrix V_m to form an orthogonal basis of $\mathcal{K}_m(A, r_0)$. According to Krylov subspace methods the approximate solution have the form

$$x_m = x_0 + V_m y_m \quad \text{for some } y_m \in R^m. \tag{3.1}$$

Different Krylov subspace methods seek $z_m = V_m y_m$ to satisfy different conditions. The Arnoldi process produces an $m \times m$ upper Hessenberg matrix H_m. Key properties of this process are that $\beta V_m e_1 = r_0$ and that the matrix H_m satisfies the relation

$$AV_m = V_{m+1}\widetilde{H}_m = V_m H_m + v_{m+1} h_{m+1,m} e_m^T, \tag{3.2}$$

where e_i denotes the i-th column of the identity matrix.

We now derive the TGMBACK algorithm. Suppose that x_0 is an initial solution estimate and that the approximate x_m^{TGB} has the form $x_m^{TGB} = x_0 + V_m y_m$, where the columns of V_m form an orthogonal basis for the Krylov subspace $\mathcal{K}_m(A, r_0)$ and y_m is a vector in R^m determined later. Using Theorem 2.2 we can write all peturbations $[\Delta, \delta]$ which satisfy

$$(A - \Delta)(x_0 + V_m y_m) = b + \delta \tag{3.3}$$

as follows (cf.(2.9))

$$[\Delta, \delta] = -r_m \frac{w_m^T}{(1 + \|x_0 + V_m y_m\|_2^2)^{1/2}} + Y(I - w_m w_m^T), \tag{3.4}$$

where

$$\begin{aligned} w_m &= \begin{pmatrix} x_0 + V_m y_m \\ 1 \end{pmatrix} / (1 + \|x_0 + V_m y_m\|_2^2)^{1/2}, \\ r_m &= b - A(x_0 + V_m y_m). \end{aligned} \tag{3.5}$$

Using (3.2) we deduce that

$$r_m = \beta v_1 - V_{m+1}\widetilde{H}_m y_m = V_{m+1}(\beta e_1 - \widetilde{H}_m y_m).$$

Instituting it into (3.4) we have

$$[\Delta, \delta] = V_{m+1}(\widetilde{H}_m y_m - \beta e_1) \frac{w_m^T}{(1 + \|x_0 + V_m y_m\|_2^2)^{1/2}} + Y(I - w_m w_m^T). \tag{3.6}$$

The TGMBACK algorithm seeks $x_m = x_0 + V_m y_m$ such that x_m minimizes the normwise backward perturbation on both the matrix A and vector b in (3.3). Namely, solve the following minimization problem (cf.(3.6))

$$\|[\Delta,\delta]_{\min}^{TGB}\|_F^2 = \min_{\substack{y_m\in R^m\\ Y\in R^{n\times n}}} \|V_{m+1}(\widetilde{H}_m y_m - \beta e_1)\frac{w_m^T}{(1+\|x_0+V_m y_m\|_2^2)^{1/2}} + Y(I - w_m w_m^T)\|_F^2. \tag{3.7}$$

Denote

$$f(y_m, Y) = \|V_{m+1}(\widetilde{H}_m y_m - \beta e_1)\frac{w_m^T}{(1+\|x_0+V_m y_m\|_2^2)^{1/2}} + Y(I - w_m w_m^T)\|_F^2. \tag{3.8}$$

The following theorem gives the solution to the minimization problem in (3.7).

Theorem 3.1. *Suppose that m steps of the Arnoldi process has been taken, then*

$$\min\{f(y_m, Y) : y_m \in R^m,\ Y \in R^{n\times n}\} = \lambda_{\min}(P, Q), \tag{3.9}$$

where P and Q are symmetric positive semidefinite and symmetric positive definite matrices, respectively:

$$P = [-\beta e_1, \widetilde{H}_m]^T[-\beta e_1, \widetilde{H}_m], \quad Q = \begin{pmatrix} x_0^T & 1 \\ V_m^T & 0 \end{pmatrix}\begin{pmatrix} x_0 & V_m \\ 1 & 0 \end{pmatrix}. \tag{3.10}$$

The smallest perturbation is given by

$$\begin{aligned}
[\Delta,\delta]_{\min}^{TGB} &= V_{m+1}(\widetilde{H}_m y_m^{TGB} - \beta e_1)\frac{\begin{pmatrix} x_0 + V_m y_m^{TGB} \\ 1 \end{pmatrix}^T}{1+\|x_0+V_m y_m^{TGB}\|_2^2},\\
\|[\Delta,\delta]_{\min}^{TGB}\|_F &= \|[\Delta,\delta]_{\min}^{TGB}\|_2 = \sqrt{\lambda_{\min}(P,Q)},
\end{aligned} \tag{3.11}$$

with the associated eigenvector $\widetilde{v} = [\widetilde{v}_1, \cdots, \widetilde{v}_{m+1}]^T$. If $\widetilde{v}_1 \neq 0$, then

$$y_m^{TGB} = [\widetilde{v}_2/\widetilde{v}_1, \cdots, \widetilde{v}_{m+1}/\widetilde{v}_1]^T.$$

Proof. Using proposition 2.1 we rewrite (3.8) as follows:

$$\begin{aligned}
f(y_m, Y) = &\Big\|\mathrm{vec}\Big(V_{m+1}(\widetilde{H}_m y_m - \beta e_1)\frac{w_m^T}{(1+\|x_0+V_m y_m\|_2^2)^{1/2}}\Big)\\
&+ [(I - w_m w_m^T)\otimes I]\mathrm{vec}(Y)\Big\|_2^2.
\end{aligned}$$

It is easy to see

$$\begin{aligned}
\nabla_{vec(Y)} f(y_m, Y) = &2[(I - w_m w_m^T)\otimes I]vec\Big(V_{m+1}(\widetilde{H}_m y_m - \beta e_1)\\
&\frac{w_m^T}{(1+\|x_0+V_m y_m\|_2^2)^{1/2}}\Big) + 2[(I - w_m w_m^T)\otimes I]vec(Y)
\end{aligned}$$

$$=2V_{m+1}(\widetilde{H}_m y_m - \beta e_1)\frac{w_m^T(I - w_m w_m^T)}{(1 + \|x_0 + V_m y_m\|_2^2)^{1/2}} + 2Y(I - w_m w_m^T).$$

The $\bigtriangledown_{vec(Y)} f(y_m, Y) = 0$ leads to

$$Y(I - w_m w_m^T) = 0.$$

Thus, the minimization of $f(y_m, Y)$ may be modified to

$$\begin{aligned}
&\min_{y_m} \left\| V_{m+1}(\widetilde{H}_m y_m - \beta e_1)\frac{[(x_0 + V_m y_m)^T, 1]^T}{1 + \|x_0 + V_m y_m\|_2^2} \right\|_F^2 \\
&= \min_{y_m} \|\widetilde{H}_m y_m - \beta e_1\|_2^2/(1 + \|x_0 + V_m y_m\|_2^2) \\
&= \min_{y_m} \frac{\left\|(-\beta e_1, \widetilde{H}_m)\begin{pmatrix} 1 \\ y_m \end{pmatrix}\right\|_2^2}{\left\|\begin{pmatrix} x_0 & V_m \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ y_m \end{pmatrix}\right\|_2^2} = \min_{y_m} \frac{\begin{pmatrix} 1 \\ y_m \end{pmatrix}^T P \begin{pmatrix} 1 \\ y_m \end{pmatrix}}{\begin{pmatrix} 1 \\ y_m \end{pmatrix}^T Q \begin{pmatrix} 1 \\ y_m \end{pmatrix}}.
\end{aligned} \tag{3.12}$$

From Courant-Fischer's theorem[13] and if in the eigenvalue problem $Pv = \lambda Qv$, the eigenspace span(V) associated with the smallest eigenvalues is not orthogonal to span(e_1). Then all condition of Theorem 3.1 follows, thus completing the proof.

Since P and Q are symmetric positive semi-definite and symmetric positive definite matrices, respectively. The eigenvalue problem

$$Pv = \lambda Qv \tag{3.13}$$

is a regular generalized symmetric eigenvalue problem[3] which never encounters degenerate eigenvectors.

From Theorem 3.1 we have the following consequences.

Corollary 3.1. *Suppose that the eigenvalues of the generalized eigenvalue problem (3.13) are arranged in increasing order:* $\lambda_1 = \lambda_2 = \cdots = \lambda_k < \lambda_{k+1} \le \cdots \le \lambda_{m+1}$, *then*

$$[\Delta, \delta]_{\min} = V_{m+1}(-\beta e_1, \widetilde{H}_m)\begin{pmatrix} \eta \\ \widetilde{y}_m \end{pmatrix} \frac{\begin{pmatrix} \eta \\ \widetilde{y}_m \end{pmatrix}^T \begin{pmatrix} x_0^T & 1 \\ V_m^T & 0 \end{pmatrix}}{\eta^2 + \|\eta x_0 + V_m \widetilde{y}_m\|_2^2}, \tag{3.14}$$

where $\begin{pmatrix} \eta \\ \widetilde{y}_m \end{pmatrix} \in span(v_1, \cdots, v_k)$, *while* v_i *is the eigenvector associated with* λ_i, $i = 1, 2, \ldots, m+1$. *Thus, if* $span(v_1, \cdots, v_k)$ *is not orthogonal to* $span(e_1)$, *then the TGMBACK solution exists: we can get* $\begin{pmatrix} \eta \\ \widetilde{y}_m \end{pmatrix} \in span(v_1, \cdots, v_k)$ *such that* $\eta \ne 0$, *then* $y_m^{TGB} = \widetilde{y}_m/\eta$. *Furthermore, if* $k > 1$, *then the solution may not be unique. If* $span(v_1, \cdots, v_k)$ *is orthogonal to* $span(e_1)$, *then the TGMBACK solution does not exist.*

Corollary 3.2. *If* P *is singular, i.e.* $h_{m+1,m} = 0$, *then* $y_m^{TGB} = \beta H_m^{-1} e_1$ *and* $x_m = x_0 + V_m y_m^{TGB}$ *is the exact solution.*

Suppose P is nonsingular, Let

$$\widehat{H}_m = [-\beta e_1, \widetilde{H}_m], \tag{3.15}$$

then $\widehat{H}_m \in R^{(m+1)\times(m+1)}$ is a nonsingular upper triangular matrix. The symmetric positive definite matrix Q can be factorized as follows:

$$Q = \begin{pmatrix} \|x_0\|_2^2+1 & x_0^T V_m \\ V_m^T x_0 & I_m \end{pmatrix} = \begin{pmatrix} 1 & \\ V_m^T x_0/(\|x_0\|_2^2+1) & I_m \end{pmatrix}$$
$$\begin{pmatrix} \|x_0\|_2^2+1 & \\ & V_m^T\Big(I_m - \dfrac{x_0x_0^T}{1+\|x_0\|_2^2}\Big)V_m \end{pmatrix} \begin{pmatrix} 1 & x_0^T V_m/(\|x_0\|_2^2+1) \\ & I_m \end{pmatrix}. \tag{3.16}$$

Obviously, the generalized eigenvalue problem (3.13) can be reduced to the following standard eigenvalue problem:

$$\widehat{H}_m^{-T} \begin{pmatrix} 1 & \\ V_m^T x_0/(\|x_0\|_2^2+1) & I_m \end{pmatrix} \begin{pmatrix} \|x_0\|_2^2+1 & \\ & V_m^T\Big(I_m - \dfrac{x_0x_0^T}{1+\|x_0\|_2^2}\Big)V_m \end{pmatrix}$$
$$\begin{pmatrix} 1 & x_0^T V_m/(\|x_0\|_2^2+1) \\ & I_m \end{pmatrix} \widehat{H}_m^{-1} z = \xi z, \tag{3.17}$$

where $z = \widehat{H}_m v$ and $\xi = 1/\lambda$. From (3.17) it is easy to see that the eigenvalue ξ and eigenvector z of (3.17) can be computed from the singular value $\sqrt{\xi}$ and the associated right singular vector z by the singular value decomposition of the following matrix:

$$\begin{pmatrix} (\|x_0\|_2^2+1)^{1/2} & x_0^T V_m/(x_0\|_2^2+1)^{1/2} \\ 0 & \Big[V_m^T\Big(I_m - \dfrac{x_0x_0^T}{1+\|x_0\|_2^2}\Big)V_m\Big]^{1/2} \end{pmatrix} \widehat{H}_m^{-1}. \tag{3.18}$$

4. Implementation

In order to simplify the eigenvalue problem, we change the express for the minimization of $f(y_m, Y)$ (cf.(3.12)) as follows

$$\min_{y_m} \|\widetilde{H}_m y_m - \beta e_1\|_2^2/(1+\|x_0+V_m y_m\|_2^2) = \min_{y_m} \frac{\|(\widetilde{H}_m, -\beta e_1)\begin{pmatrix} y_m \\ 1 \end{pmatrix}\|_2^2}{\|\begin{pmatrix} V_m & x_0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} y_m \\ 1 \end{pmatrix}\|_2^2}$$
$$= \min_{y_m} \frac{\begin{pmatrix} y_m \\ 1 \end{pmatrix}^T \widetilde{P} \begin{pmatrix} y_m \\ 1 \end{pmatrix}}{\begin{pmatrix} y_m \\ 1 \end{pmatrix}^T \widetilde{Q} \begin{pmatrix} y_m \\ 1 \end{pmatrix}}, \tag{4.1}$$

where

$$\widetilde{P} = \begin{pmatrix} \widetilde{H}_m^T \widetilde{H}_m & -\beta \widetilde{H}_m^T e_1 \\ -\beta e_1^T \widetilde{H}_m & \beta^2 \end{pmatrix}, \quad \widetilde{Q} = \begin{pmatrix} I_m & V_m^T x_0 \\ x_0^T V_m & 1+\|x_0\|_2^2 \end{pmatrix}. \tag{4.2}$$

If we make the Cholesky factorization $\widetilde{Q} = LL^T$, then the lower triangular matrix L is as follows

$$L = \begin{pmatrix} I_m & 0 \\ x_0^T V_m & \sqrt{1+\|x_0\|_2^2 - \|V_m^T x_0\|_2^2} \end{pmatrix}. \tag{4.3}$$

The generalized eigenvalue problem $\widetilde{P}v = \lambda\widetilde{Q}v$ can be reduced to the following standard symmetric eigenvalue problem

$$L^{-1}\widetilde{P}L^{-T}u = \lambda u, \tag{4.4}$$

where $u = L^T v$.

Since

$$\widetilde{P} = [\widetilde{H}_m, -\beta e_1]^T[\widetilde{H}_m, -\beta e_1],$$

it is difficult to compute the smallest eigenvalue $\lambda_{\min}$ and the corresponding eigenvector of matrix $L^{-1}\widetilde{P}L^{-T}$ in (4.4) accurately, when $\sqrt{\lambda}$ is small. Thus, we compute instead the smallest singular value and the corresponding right singular vector of the following upper Hessenberg matrix:

$$[\widetilde{H}_m, -\beta e_1]L^{-T} = (\widetilde{H}_m, -\beta e_1)\begin{pmatrix} I_m & -(V_m^T x_0)tl \\ & tl \end{pmatrix} = [\widetilde{H}_m, x_t], \tag{4.5}$$

where $x_t = -tl(\widetilde{H}_m V_m^T x_0 + \beta e_1)$ and $tl = 1/\sqrt{1 + \|x_0\|_2^2 - \|V_m^T x_0\|_2^2}$.

Finally, we give the proposed algorithm as follows.

Restarted TGMBACK Algorithm: TGMBACK(m)

1. Initialize: Choose x_0, compute $r_0 = b - Ax_0$ and set $\beta = \|r_0\|_2, v_1 = r_0/\beta$.
2. The Arnoldi process

   for $j = 1, 2, \ldots, m$

   $\quad w := Av_j$

   $\quad$ for $i = 1, 2, \ldots, j$

   $\qquad h_{i,j} = (w, v_i)$

   $\qquad w := w - h_{i,j}v_i$

   $\quad h_{j+1,j} = \|w\|_2$

   $\quad v_{j+1} = w/h_{j+1,j}$

3. Compute the smallest singular value σ and the associated right singular vector u of the upper Hessenberg matrix $[\widetilde{H}_m, x_t]$ (cf.(4.5)).

   Compute $v = L^{-T}u \equiv \begin{pmatrix} \widetilde{y}_m \\ \eta \end{pmatrix}$

   Normarize the vector v to get y_m: $y_m = \widetilde{y}_m/\eta$

   Form the approximate solution $x_m^{TGB} = x_0 + V_m y_m$

4. Set $\|[\Delta, \delta]_{min}^{TGB}\|_F = \sigma$. If satisfied, stop; else set $x_0 := x_m^{TGB}$, Compute $r_0 = b - Ax_0$, set $\beta = \|r_0\|_2, v_1 = r_0/\beta$ and go to step 2.

Remark 4.1. From the proof of Theorem 3.1 or from Theorem 2.1 we have

$$\|[\Delta_m, \delta_m]_{\min}^{TGB}\|_F \equiv \sigma_m = \frac{\|r_m\|_2}{\sqrt{1 + \|x_m\|_2^2}}, \tag{4.6}$$

which can be used for checking on the correctness of the computation.

In order to retain the simple form of the matrix in (4.5), we use the left preconditioner to construct the preconditioned TGMBACK algorithm (cf. section 5).

5. Numerical Examples

A main example comes from the discretization of the convection-diffusion equation[4,11].

$$-\frac{\partial^2 u}{\partial x^2}-\frac{\partial^2 u}{\partial y^2}+\gamma\Big(x\frac{\partial u}{\partial x}+y\frac{\partial u}{\partial y}\Big)+\beta u=f,\ \text{on } (0,1)\times(0,1) \tag{5.1}$$

with zero Dirichlet boundary condition, where $\gamma = 1000$, $\beta = 10$. We discretize (5.1) using centred differences on a uniform 32×32 grid. The right-hand side was chosen such that the vector of all ones is the exact solution of the linear system. Initial guess $x_0 = 0$. As stopping criterion, we used

$$\|[\Delta, \delta]_{\min}^{TGB}\|_F \le 10^{-10}. \tag{5.2}$$

Remark 5.1. If the parameters in (5.1) are chosen to be $\beta = -200$ and $\gamma = 100$[4] or, generally, two parameters β and γ are chosen to satisfy $\beta = -2\gamma$ with $\gamma \ge 100$, then the resulting linear system is so ill conditioned that even if

$$\|[\Delta_m, \delta_m]_{\min}^{TGB}\|_F \le 10^{-12},$$

the solution x_m is far from the exact solution.

We compare the following algorithms:

(1) TGMBACK(m);

(2) GMRES(m);

(3) PTGMBACK(m,l): A preconditioned TGMBACK(m), the preconditioner $\widetilde{\Delta}(l)$ is an approximation of the disceate Laplancian difference operator $\widetilde{\Delta}$ with zero boundary condition. i.e., for a vector v given, $\widetilde{\Delta}(l)^{-1}v$ is the l-th Gauss-Seidel iterative vector $x^{(l)}$ of the following difference equation:

$$\widetilde{\Delta}x = v,$$

with initial iterative vecor $x^{(0)} = 0$.

Fig.5.1. $m = 25, l = 1$, compare minimum backward pertubation.
(i) TGMBACK(m), (ii) GMRES(m), (iii) PTGMBACK(m,l), (iv) PGMRES(m,l).

(4) PGMRES(m, l): A preconditioned GMRES(m) with the same preconditioner as PTGMBACK(m, l).

We compare the norms of the residual vectors and the norms of the minimum backward perturbations. The norm of the minimum backward perturbation for GMRES(m) and PGMRES(m) can be computed by using the residual norm $\|r_m\|_2$ and the norm of the iterative vctor $\|x_m\|_2$ (cf.(2.7))

$$\|[\Delta, \delta]_{\min}^{\mathrm{TGB}}\|_F = \|[\Delta, \delta]_{\min}^{\mathrm{TGB}}\|_2 = \|r_m\|_2 / \sqrt{1 + \|x_m\|_2^2}.$$

The results of the computation are shown in Fig.5.1–Fig.5.4.

Fig.5.2. $m = 25, l = 1$, compare residual.
(i) TGMBACK(m), (ii) GMRES(m), (iii) PTGMBACK(m, l), (iv) PGMRES(m, l).

Fig.5.3. $m = 15, l = 1$ compare minimum backward perturbation.
(i) TGMBACK(m), (ii) GMRES(m), (iii) PTGMBACK(m, l), (iv) PGMRES(m, l).

Fig.5.4. $m = 15, l = 1$ compare residual.
(i) TGMBACK(m), (ii) GMRES(m), (iii) PTGMBACK(m, l), (iv) PGMRES(m, l).

References

[1] M. Arioli, I. Duff, D. Ruiz. Stopping criteria for iterative solvers, *SIAM J. Matrix Anal. Appl.* **13** (1992), 138–144.

[2] A. Ben-Israel, T.N.E. Greville, Generalized Inverses: Theory and Applications, John Wiley & Sons, New York, 1974.

[3] Z.H. Cao, On a deflation method for the symmetric generalized eigenvalue problem, *Linear Algebra Appl.,* **92** (1987), 187–196.

[4] R.W. Freund, A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems, *SIAM J. Sci. Comput.,* **14** (1993), 470–482.

[5] G.H. Golub, C.F.Van Loan, An analysis of the total least squares problem, *SIAM J. Numer. Anal.,* **17** (1980), 883–893.

[6] G.H. Gutknecht, Variants of BICGSTAB for matrices with complex spectrum, *SIAM J. Sci. Comput.,* **14** (1993), 1020–1033.

[7] E.M. Kasenally, GMBACK: A generalized minimum backward error algorithm for nonsymmetric linear systems, *SIAM J. Sci. Comput.,* **16** (1995), 698–719.

[8] P. Lancaster, M. Tismenetsky, The Theory of Matrices, Academic Press, Orlando, Florida, 1985.

[9] W. Oettli, W. Prager, Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides, *Numer. Math.,* **6** (1964), 405–409.

[10] J.L. Rigal, J. Gaches, On the compatibility of a given solution with the data of a linear system, *J. Assoc. Comput. Mech.,* **14** (1967), 543–548.

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

[12] Y. Saad, M.H. Schultz, GMRES: A generalized minimum residual algorithm for solving non-symmetric systems, *SIAM J. Sci. Statist. Comput.,* **7** (1986), 856–869.

[13] J.H. Wilkinson, The algebraic eigenvalue problem, Clarendon Press, Oxford, 1965.