UC-405

SAND95-8220
Unlimited Release
Printed April 1995

A Modification to the GMRES Method for Ill-Conditioned Linear Systems

J. C. Meza
Scientific Computing Department
Sandia National Laboratories
Livermore, CA 94551-0969
meza@ca.sandia.gov

ABSTRACT

This paper concerns the use of a method for the solution of ill-conditioned linear systems. We show that the Generalized Minimum Residual Method (GMRES) in conjunction with a truncated singular value decomposition can be used to solve large nonsymmetric linear systems of equations which are nearly singular. Error bounds are given for the right singular vectors and singular values computed. A consequence of the error bounds results in a method for computing some of the singular values and right singular vectors for large matrices.

1. Introduction. Many important problems in numerical analysis require the solution of nearly singular linear systems. For example, in the solution of large-scale partial differential equations one often has to solve nearly singular linear systems [8, 13]. Other examples include constrained optimization problems where the constraints may be nearly linearly dependent [6], decomposable Markov chains [14], and integral equations [10].

In this study we address the issues of computing the solution of large, highly ill-conditioned linear systems of equations by using an iterative technique that computes the solution in the space spanned by the orthogonal complement of the singular vectors corresponding to the small singular values. In particular, we propose a modification to the GMRES method [12] for ill-conditioned systems of linear equations. Brown and Walker [2] have also addressed the issue of using GMRES for nearly singular systems and have suggested a technique based on incremental condition estimation. Our approach also has the advantage of producing good approximations to some of the large and small singular values of the matrix as well as the corresponding right singular vectors.

Consider the system of linear equations

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

where x and b are n dimensional vectors and A is an $n \times n$ real matrix. Let the singular value decomposition (SVD) for A be

$$A = U\Sigma V^T,$$

where $U = [u_1, \ldots, u_n]$ and $V = [v_1, \ldots, v_n]$ are orthogonal matrices and $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$ such that

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n \geq 0.$$

If A is nonsingular, then the solution to (1.1) can be written in terms of the SVD as follows:

$$x = \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i.$$

In this paper, we are interested in the case where the matrix A is nearly singular, that is, where one or more of the singular values is small. For purposes of exposition, we

will only consider systems where there is one small singular value although the general case can also be handled. If there is only one small singular value, then we can split the solution to (1.1) into two components:

$$x = x_d + \frac{u_n^T b}{\sigma_n} v_n, \tag{1.2}$$

where

$$x_d = \sum_{i=1}^{n-1} \frac{u_i^T b}{\sigma_i} v_i. \tag{1.3}$$

Equation (1.2) is called the *deflated decomposition* and the vector x_d is called the *deflated solution* to (1.1). There are many definitions of the deflated solution. Chan [3] defines deflated solutions of (1.1) as solutions to nearby singular but consistent systems derived from (1.1). For example, one might choose the nearest singular matrix to A in the Frobenius norm and pick the deflated solution to be the one with minimum norm. It is well known that this choice amounts to setting the smallest singular value in the singular value decomposition of the matrix A equal to zero. Other definitions can be found in [3].

In certain applications [4] it is preferable to compute the deflated decomposition (1.2) for accuracy reasons, whereas in other applications the deflated solution is the only solution of interest. Notice that if the singular vectors u_n and v_n were known then both (1.2) and (1.3) could be computed by first computing x and then orthogonalizing against v_n. However, even if this decomposition were known this approach is not advisable because it usually results in a poor approximation to x_d due to roundoff errors. In particular, if the component of the solution in the direction of v_n is large, then errors in that component tend to dominate the solution in the other directions.

Stewart [15] suggested a method for computing the deflated solution of (1.1) by an implicit method. This method uses orthogonal projections constructed from approximations to the singular vectors of the matrix A corresponding to the smallest singular value. The disadvantage of this method is that it requires a direct method for the solution of (1.1). Chan and Saad [5] proposed a deflated Lanczos method for symmetric positive definite linear systems which only requires a matrix-vector product. This work has also been extended to nonsymmetric systems [9]. In this study, we propose a new variation which appears to be more robust than the ones studied in [9].

2. GMRES Method. Saad and Schultz [12] proposed the GMRES method for solving large sparse nonsymmetric linear systems based on the Arnoldi process [1] for computing the eigenvalues of a matrix. Arnoldi's method is just the Gram-Schmidt method for computing an orthonormal basis for a particular Krylov subspace. A statement of the GMRES algorithm is given in Algorithm 2.1.

ALGORITHM 2.1. *GMRES Method*

1. *Choose* x_0 *and compute* $r_0 = b - Ax_0$. *Set* $w_1 = r_0/\|r_0\|$.
2. *For* $j = 1, 2, ...m$

$$\begin{aligned} h_{ij} &= (Aw_j, w_i), \qquad i = 1, 2, ..., j \\ \hat{w}_{j+1} &= Aw_j - \sum_{i=1}^{j} h_{ij} w_i \\ h_{j+1,j} &= \|\hat{w}_{j+1}\| \\ w_{j+1} &= \hat{w}_{j+1}/h_{j+1,j} \end{aligned}$$

3. *Form the solution:*

$$\min ||\overline{H}_m y_m - \beta_m e_1||, \beta_m = ||r_m|| \tag{2.4}$$

$$x_m = x_0 + W_m y_m. \tag{2.5}$$

The matrix $W_m = [w_1, w_2, \ldots, w_m]$, and the entries of the $(m+1) \times m$ upper Hessenberg matrix $\overline{H}_m$ are the scalars, $h_{ij}, i = 1, \ldots, m+1; j = 1, \ldots, m$, generated in step 2 of the GMRES algorithm. It is easy to show that

$$AW_m = W_m H_m + \hat{w}_{m+1} e_m^T, \tag{2.6}$$

where the upper Hessenberg matrix H_m is the $m \times m$ matrix constructed by deleting the last row of $\overline{H}_m$.

The number of iterations, m, in step 2 of Algorithm 2.1 is chosen so that the approximate solution x_m is sufficiently accurate, but small enough so as not to be prohibitive in terms of storage required. If after m iterations the approximate solution has not converged then it is possible to restart the algorithm using the current estimate of x as the new initial guess. This method is denoted by GMRES(m), or the restarted

GMRES. The residual at any iteration may be computed without actually solving (1.1) through the relation [11],

$$||b - Ax_m|| = h_{m+1,m}|e_m^T y_m|. \tag{2.7}$$

Although the computation of the residual by (2.7) requires solving (2.4) for y_m there are ways to circumvent this computation by carrying an LU or QR factorization of the matrix H throughout the Arnoldi process.

It is well known that the Arnoldi process may be viewed as a Galerkin process for estimating the eigenvalues of a matrix. In particular, if we apply Algorithm 2.1 to a linear system of size n then the upper Hessenberg matrix, H_n, that is generated after n steps of the Arnoldi process will have the same eigenvalues as the matrix A. We might expect then that if the original matrix A is ill-conditioned that the intermediate matrices, H_m, generated by the GMRES algorithm might also be ill-conditioned. Therefore if we solve (2.4) for y_m in the straightforward way, our computed solution will be inaccurate for the reasons indicated in Section 1.

Fortunately, computing the deflated solution of (2.4) is easier than computing the deflated solution of (1.1). Since the matrix H_m has dimension $m \ll n$ the solution of (2.4) is at least computationally easier. Moreover, the matrix elements of H_m are on hand whereas the matrix elements of A may not be available, as for example in the inner iteration of a nonlinear method. The next section describes one such technique which can be used to compute the deflated solution of (2.4) in a stable manner.

Several deflation techniques have been previously suggested for the solution of nearly singular nonsymmetric linear systems [9]. We propose a new method based on the truncated SVD solution to (2.4). Assume that the matrix $\overline{H}_m$ has exactly one small singular value and consider its singular value decomposition,

$$\overline{H}_m = U\Theta V^T,$$

where U and V are orthogonal matrices and Θ is a diagonal matrix containing the singular values of $\overline{H}_m$. By a truncated least squares solution to the system

$$\overline{H}_m y_m = f,$$

we will mean the solution obtained by setting $\theta_m = 0$, so that the solution is given by

$$y_m = \sum_{i=1}^{m-1} \frac{u_i^T f}{\theta_i} v_i.$$

3. Theoretical Results. A question of when the singular values and singular vectors of the matrix $\overline{H}_m$ converge to the singular values and vectors of the matrix A still remains. If in the Arnoldi process the smallest singular value of $\overline{H}_m$ is not a good approximation to the smallest singular value of the matrix A then we should not compute the deflated solution. In this section we present some results that attempt to address these issues. We will show that given an approximation to the singular vectors and a corresponding approximate singular value, we can derive some useful error bounds that can be used to determine when to deflate the solution in the computation of the truncated least squares.

Recall that the GMRES method is based on the Arnoldi method for reducing an $n \times n$ matrix to upper Hessenberg form. If the Arnoldi method were to be carried out for n steps, then in exact arithmetic we would have

$$A = W H_n W^T, \tag{3.8}$$

where W is an $n \times n$ orthogonal matrix. If we were to then compute the SVD of the upper Hessenberg matrix H_n we would have the SVD of the matrix A by rewriting (3.8) as

$$A = W U_n \Theta_n V_n^T W^T, \tag{3.9}$$

$$= Z_n \Theta_n Y_n^T, \tag{3.10}$$

where Y_n and Z_n are orthogonal matrices. Of course this is impractical, since we would never take n steps of the Arnoldi method, but this does imply that there is a relation between the intermediate matrices generated in the Arnoldi process and the singular vectors of the matrix A.

THEOREM 3.1. *Suppose that m steps of the Arnoldi method have been taken, so that we have*

$$A W_m = W_m H_m + r_m e_m^T, \tag{3.11}$$

where

$$r_m = h_{m+1,m} w_{m+1}. \tag{3.12}$$

Furthermore, let $H_m = U_m \Theta_m V_m^T$ *be the singular value decomposition of* H_m *and define*

$$Y_m = [y_1, y_2, \ldots, y_m] = W_m V_m, \tag{3.13}$$

$$Z_m = [z_1, z_2, \ldots, z_m] = W_m U_m, \tag{3.14}$$

$$\beta_m = ||r_m||. \tag{3.15}$$

Then

$$||Ay_i - \theta_i z_i||_2 \leq \beta_m |v_{mi}|, \quad i = 1, \ldots, m, \tag{3.16}$$

Proof. Multiply (3.11) on the right by V_m which gives

$$AW_m V_m = W_m H_m V_m + r_m e_m^T V_m.$$

Using the definition of Y_m, Z_m, and the SVD of H_m we have

$$AY_m = Z_m \Theta_m + r_m e_m^T V_m,$$

or component-wise

$$Ay_i = \theta_i z_i + r_m (e_m^T V_m e_i), \quad i = 1, \ldots, m.$$

Inequality (3.16) follows by taking norms and using the definition of β_m. □

It would be satisfying to have an equivalent relation for the transpose of inequality (3.16), that is,

$$A^T z_i - \theta_i y_i \leq \eta, \quad i = 1, \ldots, m. \tag{3.17}$$

Unfortunately inequality (3.17) does not hold except in the case where $m = n$ and using exact arithmetic. The best that can be achieved is

$${W_m}^T (A^T z_i - \theta_i y_i) = 0, \quad i = 1, \ldots, m. \tag{3.18}$$

Since the GMRES method does not ever use A^T this should come as no surprise.

The vectors y_i and the scalars θ_i can still be thought of as approximate right singular vectors and singular values of the matrix A; but the vectors z_i cannot be used as such.

Based on Theorem 1 it is straightforward to show that the following error bound holds for the singular values of the upper Hessenberg matrix H_m.

THEOREM 3.2. *Suppose that m steps of the Arnoldi method have been taken, and let Θ_m, Y_m, and Z_m be defined as above. If*

$$Ay_i = \theta_i z_i + \eta_1, \tag{3.19}$$

$$A^T z_i = \theta_i y_i + \eta_2. \tag{3.20}$$

Then

$$\min_{\sigma_i \in \sigma(A)} |\sigma_i - \theta| \leq \frac{\beta_m}{\sqrt{2}} \, |v_{mi}| + ||\eta_2|| \quad i = 1, \ldots, n.$$

Proof. Consider the symmetric matrix B and vectors ψ_i and η defined by:

$$B = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}, \quad \psi_i = \frac{1}{\sqrt{2}} \begin{bmatrix} y_i \\ z_i \end{bmatrix}, \quad \eta = \begin{bmatrix} \eta_2 \\ \eta_1 \end{bmatrix}.$$

Equations (3.19-3.20) can be written as

$$\begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} y_i \\ z_i \end{bmatrix} = \theta_i \begin{bmatrix} y_i \\ z_i \end{bmatrix} + \begin{bmatrix} \eta_2 \\ \eta_1 \end{bmatrix},$$

or more compactly

$$B\psi_i = \theta_i \psi_i + \eta/\sqrt{2}.$$

The result follows from a standard error bound for approximate eigenvalues and eigenvectors for symmetric matrices (see for example [7],pp. 414) and the fact that the eigenvalues of the matrix B are $\pm$ the singular values of the matrix A. □

Remark. As stated above, the theorems hold for H_m when we would really like an equivalent result for $\overline{H}_m$ since the GMRES method uses $\overline{H}_m$ to solve the least squares problem (2.4). We conjecture that the error bounds hold for $\overline{H}_m$ and the numerical results certainly point in this direction, but we have not been able to prove so.

One could also make this argument based on several properties relating the singular values of the Hessenberg matrices generated at step m and step $m+1$. Let H_m, H_{m+1},

be the upper Hessenberg matrices generated after m and $m+1$ steps respectively of the GMRES process. The matrices $\overline{H}_m, \overline{H}_{m+1}$ are defined similarly. Then the following properties hold:

1. The singular values of $\overline{H}_{m+1}$ interlace the singular values of $\overline{H}_m$

$$\sigma_i(\overline{H}_{m+1}) \geq \sigma_i(\overline{H}_m) \geq \sigma_{i+1}(\overline{H}_{m+1}), i = 1, \ldots m. \tag{3.21}$$

2. The singular values of H_{m+1} interlace the singular values of $\overline{H}_m$,that is,

$$\sigma_i(H_{m+1}) \geq \sigma_i(\overline{H}_m) \geq \sigma_{i+1}(H_{m+1}), i = 1, \ldots m. \tag{3.22}$$

3. The singular values of H_m and $\overline{H}_m$ are related by

$$\sigma_i(H_m) \leq \sigma_i(\overline{H}_m), \quad i = 1, \ldots, m \tag{3.23}$$

Of these properties, the first one is the most relevant for our purposes. This property tells us that the smallest singular value is a nonincreasing function of the step m. Therefore if the smallest singular value of $\overline{H}_m$ is ever small then we know that the matrix A must have at least one singular value that is at least as small as the one computed from the singular value decomposition of $\overline{H}_m$.

4. Numerical Results. This section presents several numerical experiments comparing the various methods described in Section 3.

Recall that the linear system of interest is

$$Ax = b,$$

where x and b are n dimensional vectors and A is an $n \times n$ real matrix which is nearly singular. The numerical experiments were run on an SGI 4D/25, using double precision arithmetic (machine epsilon $\approx 10^{-16}$). The method was said to converge whenever

$$\|r_k\|/\|r_0\| \leq \epsilon,$$

where $\epsilon = 10^{-9}$. We will denote the new modification to the GMRES method by the term GMSVD. For comparison purposes we also tested an algorithm based on computing the solution to the linear system by the GMRES method and then computing the

deflated solution by orthogonalizing against the null vector as computed from a singular value decomposition of the matrix A. This method will be referred to as GMRESD.

To compare the two methods, we computed both the deflated error and the deflated residual. By the deflated error and deflated residual we mean the projection of the error and residual into the subspace spanned by the singular vectors corresponding to the large singular values. In our case this means the subspace spanned by all of the singular vectors except for the one corresponding to the small singular value. An easy way to compute these quantities is to define the projection operators

$$P_v = I - v_m v_m^T, \tag{4.24}$$

$$P_u = I - u_m u_m^T, \tag{4.25}$$

and the deflated error and residual by the formulas

$$e_d = P_v e = P_v(x - \hat{x}), \tag{4.26}$$

$$r_d = P_u r = P_u(b - A\hat{x}), \tag{4.27}$$

where $\hat{x}$ is the computed solution.

Test case 1. The first test case consists of taking a symmetric positive definite matrix and perturbing it by adding a small nonsymmetric term. Define $A(\eta) = D + \eta E$, where the matrix D is defined by $D = diag\ (10^{-J}, 2, 3, \ldots, n)$, and $J = 1, 2, \ldots, 10$. The matrices, E, are computed by generating uniform random numbers between [-0.5, +0.5], and normalizing so that $\|E\|_2 = 1$. The amount of nonsymmetry can then be adjusted by varying the noise level η. For this test problem, we chose $n = 100$ and $m = 20$.

Table 1 contains the deflated errors and deflated residuals for test case 1, using a noise level of 10^{-6}. Similar results can be obtained for larger values of η. The results clearly indicate the advantage of using the GMSVD method to compute both the deflated errors and residuals. Using the new method, both the deflated error and the deflated residual can be computed accurately irregardless of the condition number of the matrix, while the unmodified GMRES solution deteriorates as the condition number increases. In Table 2, we present the error in the computed singular values for test case 1 for various values of the noise level.

TABLE 1

Deflated errors and residuals for test case 1 (η = 1.e-6).

	Deflated errors		Deflated residuals	
J	GMRESD	GMSVD	GMRESD	GMSVD
1	1.3510E-04	1.3467E-04	1.3509E-04	1.3510E-04
2	9.5763E-05	1.1080E-04	9.5777E-05	9.5768E-05
3	2.1539E-03	6.3464E-07	2.1638E-03	1.2939E-07
4	2.1247E-02	7.7085E-07	2.0247E-02	3.9382E-07
5	5.1776E-02	8.0724E-07	2.0803E-02	3.7181E-07
6	4.8835E-01	7.8194E-07	2.0567E-02	4.1053E-07
7	2.9554E+00	7.9254E-07	2.0462E-02	4.2402E-07
8	2.7997E+01	8.3944E-07	5.3001E-03	3.8411E-07
9	2.0082E+01	8.1100E-07	1.3838E-02	3.9666E-07
10	9.4671E+00	7.9181E-07	1.6598E-02	3.7946E-07

Test case 2. The purpose of the second test problem is to simulate a typical linear system arising in a seismic inversion problem. These problems have one or more small singular values and one or more large singular values with the rest of the spectrum fairly well-conditioned. To simulate this problem, we set the matrix $D = diag\ (10^{-J}, 1, \ldots, 3, 3000)$, with the values of d_2 through d_{n-1} varying uniformly between 1 and 3. We then generate two random orthogonal matrices, Q_1 and Q_2, which are used to compute the matrix $A = Q_1 D Q_2$. This example generates a non-symmetric matrix that is well-conditioned if the small and large eigenvalues are excluded, which is typical of some of the velocity inversion problems. In this problem, we have chosen $n = 1000$ and $m = 25$.

The results for the seismic prototype problem (test case 2) are given in Table 3. The errors in the computed singular values and the norm of the difference between the right singular vector computed by GMSVD and the true right singular vector for the matrix A are given in columns 3-4 of Table 3. We note that after only 25 iterations both the smallest singular value and the corresponding right singular value of the large matrix A are very well approximated.

TABLE 2

Error in singular values for test case 1 with various levels of noise, η.

J	$\eta = 1.0^{-6}$	$\eta = 1.0^{-3}$	$\eta = 1.0^{-1}$
1	1.0000E-01	9.9909E-02	5.67E-07
2	9.9999E-03	9.9490E-03	3.04E-02
3	3.6103E-12	9.4902E-04	6.32E-07
4	3.3673E-10	4.1529E-10	2.65E-07
5	3.7942E-09	2.8625E-05	2.17E-06
6	3.6728E-08	3.1217E-09	1.42E-06
7	1.4888E-07	2.7816E-10	1.47E-06
8	2.5971E-07	6.3853E-09	2.23E-07
9	2.5525E-07	4.7789E-09	4.01E-07
10	2.2961E-07	5.2541E-09	7.76E-07

TABLE 3

Results for seismic prototype example, $n = 1000$.

J	r_d	$\lvert\sigma_n - \theta_m\rvert$	$\lvert\lvert y - v_n\rvert\rvert_2$
1	8.72E-08	2.12E-08	3.14E-05
2	4.96E-07	1.48E-07	2.86E-05
3	1.29E-06	8.53E-08	9.21E-06
4	3.27E-06	9.34E-09	9.88E-07
5	3.27E-05	9.36E-10	9.89E-08
6	3.27E-04	9.36E-11	9.89E-09
7	3.26E-03	6.28E-11	2.50E-09
8	2.80E-03	2.42E-08	2.31E-08
9	1.40E-05	3.26E-07	2.31E-07
10	1.37E-06	3.27E-06	2.31E-06

5. Conclusions. We have presented a modification of the GMRES method that can accurately compute the solution of a linear system in the presence of ill-conditioning. We have shown that the deflated residuals and errors can be computed accurately using the new method, irrespective of the condition number of the linear system. We have also

given error bounds for the approximate singular values and right singular vectors. Both of these quantities can be computed cheaply as part of the process of solving the linear system. We have also shown that the small singular values and the corresponding right singular vectors can be approximated accurately using the modified GMRES method.

Acknowledgments. The author thanks I. Olkin for his many useful suggestions and encouraging comments.

REFERENCES

[1] W. Arnoldi, *The principle of minimized iteration in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] P. Brown and H. Walker, *GMRES on (nearly) singular systems*, Tech. Report UCRL-JC-115882, Lawrence Livermore National Laboratory, Livermore, California, 1994.

[3] T. Chan, *Deflated decomposition of solutions of nearly singular systems*, SIAM Journal of Numerical Analysis, 21 (1984), pp. 738–754.

[4] ——, *Deflation techniques and block-elimination algorithms for solving bordered singular systems*, SIAM Journal of Scientific and Statistical Computing, 5 (1984), pp. 121–134.

[5] T. Chan and Y. Saad, *Deflated lanczos procedures for solving nearly singular systems*, Research Report YALEU/DCS/RR-403, Department of Computer Science, Yale University, 1985.

[6] P. Gill, W. Murray, and M. Wright, *Practical Optimization*, Academic Press, New York, 1979.

[7] G. Golub and C. van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[8] J. Meza and J. Grcar, *Dancir: A three-dimensional steady-state semiconductor device simulator*, Technical Report SAND-8266, Sandia National Laboratories, Livermore, CA, 1990.

[9] J. Meza and W. Symes, *Deflated krylov methods for nearly singular linear systems*, Journal of Optimization Theory and Applications, 72 (March, 1992), pp. 441–458.

[10] D. O'leary and J. Simmons, *A bidiagonalization-regularization procedure for large scale discretizations of ill-posed problems*, SIAM Journal of Scientific and Statistical Computing, 2 (December, 1981), pp. 474–489.

[11] Y. Saad, *Krylov subspace methods for solving large unsymmetric linear systems*, Mathematics of Computation, 37 (1981), pp. 105–126.

[12] Y. Saad and M. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal of Scientific and Statistical Computing, 7 (1986), pp. 856–869.

[13] F. Santosa and W. Symes, *An Analysis of Least Squares Velocity Inversion*, Society of Exploration Geophysicists, Tulsa, Oklahoma, 1989.

[14] G. Stewart, *Computable error bounds for aggregated markov chains*, Technical Report 901, University of Maryland Computer Science Center, College Park, MD, 1980.

[15] G. Stewart, *On the implicit deflation of nearly singular systems of linear equations*, SIAM Journal of Numerical Analysis, 2 (1981), pp. 136–140.