A stabilized GMRES method for singular and severely ill-conditioned systems of linear equations

Zeyu Liao\(^1\) \(\bullet\) Ken Hayami\(^2\) \(\bullet\) Keiichi Morikuni\(^3\) \(\bullet\) Jun-Feng Yin\(^4\)

Abstract
Consider using the right-preconditioned GMRES (AB-GMRES) for obtaining the minimum-norm solution of inconsistent underdetermined systems of linear equations. Morikuni (Ph.D. thesis, 2013) showed that for some inconsistent and ill-conditioned problems, the iterates may diverge. This is mainly because the Hessenberg matrix in the GMRES method becomes very ill-conditioned so that the backward substitution of the resulting triangular system becomes numerically unstable. We propose a stabilized GMRES based on solving the normal equations corresponding to the above triangular system using the standard Cholesky decomposition. This has the effect of shifting upwards the tiny singular values of the Hessenberg matrix which lead to an inaccurate solution. We analyze why the method works. Numerical experiments show that the proposed method is robust and efficient, not only for applying AB-GMRES to underdetermined systems, but also for applying GMRES to severely ill-conditioned range-symmetric systems of linear equations.

Keywords
Least squares problems \(\bullet\) Krylov subspace methods \(\bullet\) GMRES \(\bullet\) Inconsistent systems \(\bullet\) Minimum-norm solution \(\bullet\) Regularization

1 Introduction
As a motivating instance when the generalized minimal residual (GMRES) method iterates diverge due to severe ill-conditioning, consider obtaining the minimum-norm solution of the inconsistent least squares problem:

\[
\min_{x \in \mathbb{R}^n} ||x||_2, \text{ such that } x \in \{ \arg \min_{\xi \in \mathbb{R}^n} ||b - A\xi||_2 \}
\]
where $A \in \mathbb{R}^{m \times n}$ and $b \notin \mathcal{R}(A) \subseteq \mathbb{R}^m$. Here, $\mathcal{R}(A)$ denotes the range space of $A$. Such problems may occur in ill-posed problems where $b$ is given by an observation which contains noise. The problem (1) is equivalent to

$$(A^T A)^2 v = A^T b, \quad x = A^T A v,$$  

(2)

and the solution can be expressed by $x = A^T b$, where $A^T$ denotes the transpose of $A$ and $A^\dagger$ is the pseudoinverse of $A$. (See e.g. [2].)

The standard direct method for solving the least squares problem (1) is to use the QR decomposition. However, when $A$ is large and sparse, iterative methods become necessary. The CGLS [12] and LSQR [25] are mathematically equivalent to applying the conjugate gradient (CG) method to the normal equations of the first kind

$$A^\dagger A x = A^\dagger b,$$  

(3)

which is equivalent to

$$\min_{x \in \mathbb{R}^n} \| b - A x \|_2^2.$$  

(4)

CGLS will converge to the minimum-norm solution $x = A^\dagger b$, provided $x_0 \in \mathcal{R}(A^\dagger)$ (See, e.g. [2], p. 291). However, the convergence of these methods deteriorates for ill-conditioned problems and they require reorthogonalization [11] to improve the convergence. Here, we say (1) is ill-conditioned if the condition number $\kappa_2(A) = \|A\|_2 \|A^\dagger\|_2 \gg 1$. Alternatively, the LSMR [7] is mathematically equivalent to applying MINRES [24] to (3).

Hayami et al. [11] proposed preconditioning the $m \times n$ rectangular matrix $A$ of the least squares problem by an $n \times m$ rectangular matrix $B$ from the right and the left, and using the generalized minimal residual (GMRES) method [28] for solving the preconditioned least squares problems (AB-GMRES and BA-GMRES methods, respectively). For ill-conditioned problems, AB-GMRES and BA-GMRES were shown to be more robust compared to the preconditioned CGNE and CGLS, respectively. Note here that BA-GMRES works with Krylov subspaces in $n$-dimensional space, whereas AB-GMRES works with Krylov subspaces in $m$-dimensional space. Since $m < n$ in the underdetermined case, AB-GMRES works in a smaller dimensional space than BA-GMRES and should be more computationally efficient compared to BA-GMRES for each iteration. Moreover, AB-GMRES has the advantage that the weight of the norm in (1) does not change for arbitrary $B$. Thus, we mainly focus on using AB-GMRES to solve the underdetermined least squares problem (1). Morikuni [18] showed that AB-GMRES may fail to converge to a least squares solution in finite-precision arithmetic for inconsistent problems. We will review this phenomenon. The GMRES applied to inconsistent problems was also studied in other papers[4, 5, 19, 20, 26]. See e.g., [4, 17, 20] for methods for solving nearly singular systems.

In this paper, we first analyze the deterioration of convergence of AB-GMRES. To overcome the deterioration, we use the normal equations of the upper triangular matrix arising in AB-GMRES to change the inconsistent subproblem to a consistent one. In finite precision arithmetic, forming the normal equations for the subproblem will not
square its condition number as would be predicted by theory. In the ill-conditioned case, the tiny singular values are shifted upwards due to rounding errors. Then, applying the standard Cholesky decomposition to the normal equations will result in a well-conditioned lower triangular matrix, which will ensure that the forward and backward substitutions work stably, and overcome the problem. Our approach using the normal equations can be considered as a case where rounding errors are beneficial [14]. We analyze why the proposed method works. Numerical experiments on least squares problems with ill-conditioned rectangular coefficient matrices (Maragal\textsubscript{3T} to 7T [6]) show that the proposed method converges to a more accurate numerical solution than the original AB-GMRES. We also show by numerical experiments that the method is effective for applying GMRES to inconsistent range-symmetric systems with singular or severely ill-conditioned square coefficient matrices.

The rest of the paper is organized as follows. In Sect. 2, we briefly review AB-GMRES. In Sect. 3, we demonstrate the deterioration of convergence of AB-GMRES applied to underdetermined inconsistent least squares problems. In Sect. 4, we propose and present the stabilized GMRES method which is based on normal equations and has a regularization effect for ill-conditioned problems. We also explain why the method works by performing a rounding error analysis of the method. In Sect. 5, numerical experiment results for applying AB-GMRES to inconsistent underdetermined systems, and for applying GMRES to inconsistent systems with severely ill-conditioned and singular range-symmetric square coefficient matrices are presented. In Sect. 6, we conclude the paper.

All the experiments in this paper were done using MATLAB R2017b in double precision, unless specified otherwise (where we extended the arithmetic precision using the Multiprecision Computing Toolbox for MATLAB [1]), and the computer used was Alienware 15 CAAA015404JP with CPU Inter(R) Core(TM) i7-7820HK (2.90GHz).

2 Deterioration of convergence of AB-GMRES for inconsistent problems

In this section, we review previous work. First, we introduce the right-preconditioned GMRES (AB-GMRES). Then, we demonstrate the deterioration of convergence of AB-GMRES for inconsistent problems. Finally, we cite a related theorem to analyze the deterioration.

2.1 AB-GMRES

The AB-GMRES method of Hayami et al. [11] applies the GMRES method [28] to

\[
\min_{u \in \mathbb{R}^m} \| b - ABu \|_2, \quad x = Bu \tag{5}
\]

where \( B \in \mathbb{R}^{n \times m} \) is a preconditioning matrix.

Note the equivalence between the least squares problem (4) and the precondioned least squares problem (5).
Theorem 1 (Theorem 3.1 of [11])

\[ \min_{x \in \mathbb{R}^n} \| b - Ax \|_2 = \min_{u \in \mathbb{R}^m} \| b - ABu \|_2 \]

holds for all \( b \in \mathbb{R}^m \) if and only if \( \mathcal{R}(A) = \mathcal{R}(AB) \).

Lemma 2 (Lemma 3.3 of [11]) \( \mathcal{R}(A^T) = \mathcal{R}(B) \implies \mathcal{R}(A) = \mathcal{R}(AB) \).

Theorem 3 (Theorem 3.6 of [11]) If \( \mathcal{R}(A^T) = \mathcal{R}(B) \), then \( \mathcal{R}(AB) = \mathcal{R}(B^T A^T) \iff \mathcal{R}(A) = \mathcal{R}(B^T) \).

The convergence conditions of AB-GMRES are given as follows.

Theorem 4 (Theorem 3.7 of [11]) If \( \mathcal{R}(A^T) = \mathcal{R}(B) \), then AB-GMRES determines a least squares solution of

\[ \min_{x \in \mathbb{R}^n} \| \mathbf{b} - Ax \|_2 \]

for all \( b \in \mathbb{R}^m \) and for all \( x_0 \in \mathbb{R}^n \) if and only if \( \mathcal{R}(A) = \mathcal{R}(B^T) \).

Here, \( x_0 = Bu_0 \) is the initial approximate solution of (5) when applying AB-GMRES.

Let \( r = b - Ax = b - ABu \). Note

\[ \| r \|_2^2 = \| r \|_{\mathcal{R}(A)}^2 + \| r \|_{\mathcal{R}(A)^\perp}^2 = \| r \|_{\mathcal{R}(A)}^2 + \| b \|_{\mathcal{R}(A)^\perp}^2. \]

(6)

Here, \( S^\perp \) denotes the orthogonal complement of a subspace \( S \), and \( r \|_{\mathcal{R}(A)} \) is the \( \mathcal{R}(A) \) component of a vector \( r \). \( r \|_{\mathcal{R}(A)^\perp} \) is the \( \mathcal{R}(A)^\perp \) (inconsistent) component of the residual vector \( r \). Thus, AB-GMRES minimizes \( \| r \|_2^2 \), and hence \( \| r \|_{\mathcal{R}(A)}^2 \).

The \( k \)th iterate \( x_k \) of AB-GMRES is given by

\[ x_k = x_0 + Bu_k, \]

(7)

where \( u_k \in K_k(AB, r_0) = \text{span}\{ r_0, ABr_0, \ldots, (AB)^{k-1}r_0 \} \), so that \( x_k = x_0 + z_k \), where \( z_k \in K_k(BA, Br_0) = \text{span}\{ Br_0, (BA)Br_0, \ldots, (BA)^{k-1}Br_0 \} \). Hence, if \( x_0 \in \mathcal{R}(B) \), \( x_k \in \mathcal{R}(B) \).

If \( \mathcal{R}(B) = \mathcal{R}(A^T) \), then \( x_k \in \mathcal{R}(A^T) = \mathcal{N}(A)^\perp \). Further, if \( \mathcal{R}(B^T) = \mathcal{R}(A) \), then AB-GMRES determines a least squares solution \( x_k \), i.e., \( r_k \|_{\mathcal{R}(A)} = 0 \), where \( r_k = b - Ax_k \), and that solution \( x_k \) is the minimum Euclidean norm solution.

The algorithm is given in Algorithm 1 [11]. Here, \( H_{k+1,k} = (h_{ij}) \in \mathbb{R}^{(k+1) \times k} \) and \( e_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^{k+1} \). Algorithm 1 is said to break down when \( h_{k+1,k} = 0 \). See Appendix B of [19].
To find $y_k \in \mathbb{R}^k$ that minimizes the $k$th residual norm $\|r_k\|_2 = \|r_0\|_2 e_1 - H_{k+1,k} y_k\|_2$ in Algorithm 1, the standard approach computes the QR decomposition of $H_{k+1,k}$

$$H_{k+1,k} = Q_{k+1} R_{k+1,k}, \quad R_{k+1,k} = \begin{pmatrix} R_k \\ 0^T \end{pmatrix} \in \mathbb{R}^{(k+1) \times k},$$

(8)

where $Q_{k+1} \in \mathbb{R}^{(k+1) \times (k+1)}$ is an orthogonal matrix and $R_k \in \mathbb{R}^{k \times k}$ is an upper triangular matrix. Then, backward substitution is used to solve a system with the coefficient matrix $R_k$ as follows

$$\|r_k\|_2 = \min_{y \in \mathbb{R}^k} \|Q_{k+1}^T (\beta e_1) - R_{k+1,k} y\|_2,$$

(9)

where

$$\beta = \|r_0\|_2, \quad Q_{k+1}^T \beta e_1 = \begin{pmatrix} t_k \\ \rho_{k+1} \end{pmatrix}, \quad t_k \in \mathbb{R}^k, \quad \rho_{k+1} \in \mathbb{R}.$$  

(10)

Therefore,

$$y_k = \arg\min_{y \in \mathbb{R}^k} \|Q_{k+1}^T (\beta e_1) - R_{k+1,k} y\|_2 = R_k^{-1} t_k,$$

(11)

and the $k$th iterate is given by

$$x_k = x_0 + V_k y_k, \quad V_k = [v_1, v_2, \ldots, v_k] \in \mathbb{R}^{n \times k}, \quad V_k^T V_k = I,$$

(12)

where $I$ is the identity matrix, and $v_1, v_2, \ldots, v_k$ are the basis vectors of $K_k(AB, r_0)$ defined in Algorithm 1.

From now on, we use AB-GMRES to solve (1) with $B = A^T$ and $x_0 \in \mathcal{R}(A^T)$, e.g. $x_0 = 0$, which means $x_k = x_0 + z_k$, where $z_k \in K_k(A^T A, A^T r_0)$. Hence, Theorem 4 guarantees the convergence in exact arithmetic even in the inconsistent case. However, in finite precision arithmetic, AB-GMRES may fail to converge to a least squares solution for inconsistent problems, as shown later.
2.2 AB-GMRES for inconsistent problems

In this section, we perform experiments to show that the convergence of AB-GMRES deteriorates for inconsistent problems. Experiments were done on the transpose of the matrix Maragal_3 [6], denoted by Maragal_3T etc. Table 1 gives the information on the Maragal matrices, including the density of nonzero entries, rank and condition number. Here, the rank and condition number were determined by using the MATLAB functions spnrank [8] and svd, respectively.

Figure 1 shows the relative residual norm $\|A^T r_k\|_2 / \|A^T b\|_2$ and $\kappa_2(R_k)$ versus the number of iterations for AB-GMRES with $B = A^T$ for Maragal_3T, where $r_k = b - Ax_k$, and the vector $b$ was generated by the MATLAB function rand which returns a vector whose entries are uniformly distributed in the interval $(0, 1)$. Therefore, generically $b \notin \mathcal{R}(A)$ and the problem is inconsistent. Here, $\kappa_2(R_k) = \kappa_2(H_{k+1,k})$ holds from (8). The value of $\kappa_2(R_k)$ was computed by the MATLAB function cond. The relative residual norm $\|A^T r_k\|_2 / \|A^T b\|_2$ decreased to $10^{-8}$ until the 525th iteration, and then increased sharply. The value of cond($R_k$) started to increase rapidly around iterations 450–550. This observation shows that $R_k$ becomes ill-conditioned before convergence. Thus, AB-GMRES failed to converge to a least squares solution. This phenomenon was observed by Morikuni[18].

The reason why $R_k$ becomes ill-conditioned before convergence in the inconsistent case will be explained by a theorem in the next subsection.

Table 1 Information on the Maragal matrices

| Matrix      | $m$  | $n$  | Density [%] | Rank | $\kappa_2(A)$     |
|------------|------|------|-------------|------|-------------------|
| Maragal_3T | 858  | 1682 | 1.27        | 613  | $1.10 \times 10^3$|
| Maragal_4T | 1027 | 1964 | 1.32        | 801  | $9.33 \times 10^6$|
| Maragal_5T | 3296 | 4654 | 0.61        | 2147 | $1.19 \times 10^5$|
| Maragal_6T | 10144| 21251| 0.25        | 8331 | $2.91 \times 10^6$|
| Maragal_7T | 26525| 46845| 0.10        | 20843| $8.91 \times 10^6$|

![Fig. 1 $\kappa_2(R_k)$ and relative residual norm versus the number of iterations for Maragal_3T](image-url)
2.3 GMRES for inconsistent problems

Brown and Walker [4] analyzed the break-down of GMRES. Let $\hat{b}|_{\mathcal{R}(\hat{A})}$ denote the orthogonal projection of $b$ onto $\mathcal{R}(\hat{A})$. Assume $\mathcal{N}(\hat{A}) = \mathcal{N}(\hat{A}^T)$ and grade$(\hat{A}, \hat{b}|_{\mathcal{R}(\hat{A})}) = k$. Here, grade$(\hat{A}, \hat{b})$ for $\hat{A} \in \mathbb{R}^{n \times m}$, $\hat{b} \in \mathbb{R}^m$ is defined as the minimum $k$ such that $\mathcal{K}_{k+1}(\hat{A}, \hat{b}) = \mathcal{K}_{k}(\hat{A}, \hat{b})$. Then,

$$\dim(\mathcal{K}_k(\hat{A}, b|_{\mathcal{R}(\hat{A})})) = \dim(\mathcal{K}_{k+1}(\hat{A}, b|_{\mathcal{R}(\hat{A})}))$$

$$= \dim(\mathcal{K}_k(\hat{A}, b|_{\mathcal{R}(\hat{A})}))$$

$$= \dim(\mathcal{K}_{k+1}(\hat{A}, b|_{\mathcal{R}(\hat{A})})) = k$$

(See Appendix 1). Since $\mathcal{N}(\hat{A}) = \mathcal{N}(\hat{A}^T)$, we obtain $\hat{A}b|_{\mathcal{R}(\hat{A})} = \hat{A}b$ and $\dim(\hat{A}\mathcal{K}_{k+1}(\hat{A}, b)) = \dim(\hat{A}\mathcal{K}_{k+1}(\hat{A}, b|_{\mathcal{R}(\hat{A})})) = k$. If $b \notin \mathcal{R}(\hat{A})$ and $\dim(\hat{A}\mathcal{K}_k(\hat{A}, b)) = k$, $\dim(\hat{A}\mathcal{K}_{k+1}(\hat{A}, b)) = k + 1$ (See Appendix 2).

Let $x_0$ be the initial solution and $r_0 = b - \hat{A}x_0$. In the inconsistent case, a least squares solution is obtained at iteration $k$, and at iteration $k + 1$ breakdown occurs because of $\dim(\hat{A}\mathcal{K}_{k+1}(\hat{A}, r_0)) < \dim(\mathcal{K}_{k+1}(\hat{A}, r_0))$, i.e. rank deficiency of $\min_{r_0 \in \mathcal{K}_{k+1}(\hat{A}, r_0)} \|b - \hat{A}(x_0 + z)\|_2 = \min_{z \in \mathcal{K}_{k+1}(\hat{A}, r_0)} \|r_0 - \hat{A}z\|_2$ [4]. This case is also called the benign breakdown [26].

However, even if $\mathcal{N}(\hat{A}) = \mathcal{N}(\hat{A}^T)$, when (13) is inconsistent, the least squares problem $\min_{r_0 \in \mathcal{K}_{k+1}(\hat{A}, r_0)} \|r_0 - \hat{A}z\|_2$ may become ill-conditioned as shown below.

Brown and Walker [4] introduced an effective condition number to explain why GMRES fails to converge for inconsistent least squares problems

$$\min_{x \in \mathbb{R}^n} \|b - \hat{A}x\|_2,$$

where $\hat{A} \in \mathbb{R}^{n \times m}$ is singular, in the following Theorem 5.

**Theorem 5** [4] Assume $\mathcal{N}(\hat{A}) = \mathcal{N}(\hat{A}^T)$, and denote the least squares residual of (13) by $r^*$, the residual at the $(k - 1)$st iteration by $r_{k - 1}$. If $r_{k - 1} \neq r^*$, then

$$k_2(A_k) \geq \frac{\|A_k\|_2}{\|\hat{A}_k\|_2} \frac{\|r_{k - 1}\|_2}{\sqrt{\|r_{k - 1}\|_2^2 - \|r^*\|_2^2}},$$

(14)

where $A_k \equiv \hat{A}|_{\mathcal{K}_k(A, r_0)}$ and $\hat{A}_k \equiv \hat{A}|_{\mathcal{K}_k(A, r_0) + \text{span}(r^*)}$. Here, $\hat{A}|_S$ is the restriction of $\hat{A}$ to a subspace $S \subseteq \mathbb{R}^m$.

Theorem 5 implies that GMRES suffers ill-conditioning for $b \notin \mathcal{R}(\hat{A})$ as $\|r_k\|$ approaches $\|r^*\|$. We can apply Theorem 5 to AB-GMRES for least-squares problems by setting $\hat{A} \equiv \hat{A}\hat{A}^T$. Theorem 5 also implies that even if we choose $B$ as $A^T$, which satisfies the conditions in Theorem 4, AB-GMRES still may not converge numerically because of the ill-conditioning of $R_k$, losing accuracy in the solution computed in finite-precision arithmetic when $r_{k - 1}$ approaches $r^*$. 

© Springer
3 Deterioration of convergence of AB-GMRES applied to inconsistent underdetermined least squares problems

In this section, we illustrate the deterioration of convergence of GMRES by numerical experiments. There are two points to note in this section. The first point is that the condition number of $R_k$ tends to become very large as the iteration proceeds for inconsistent problems, as already mentioned in section 2.2. Due to $H_{k+1} = Q_{k+1}R_{k+1}$, the condition number of $H_{k+1}$ is the same as that of $R_k$, and will also become very large. The second point is as follows. Since $y_k = R_k^{-1}t_k$, $y_k$ is obtained by applying backward substitution to the triangular system

$$R_k y_k = t_k. \hspace{1cm} (15)$$

When the triangular system becomes ill-conditioned, backward substitution becomes numerically unstable, and fails to give an accurate solution $y_k$.

Figure 1 shows that at step 550 the relative residual norm suddenly increases. To understand this increase, observe the singular values of $R_{550}$.

The left of Fig. 2 shows the singular values of $R_{550}$ which were computed in double precision arithmetic. The smallest singular value of $R_{550}$ is $3.21 \times 10^{-14}$, which means that the triangular matrix $R_{550}$ is very ill-conditioned and nearly singular in double precision arithmetic.

The right of Fig. 2 shows the singular values of $R_{550}$ which were computed in quadruple precision arithmetic using the Multiprecision Computing Toolbox for MATLAB [1]. The smallest singular value of $R_{550}$ is $5.39 \times 10^{-15}$. Since quadruple precision is more accurate, from now on, we mainly show singular value distributions computed in quadruple precision.

Figure 3 shows $\kappa_2(R_k)$, $\|y_k\|_2$, and the relative residual norm $\|t_k - R_k y_k\|_2 / \|t_k\|_2$ versus the number of iterations for AB-GMRES. The relative residual norm increases only gradually when the condition number of $R_k$ is less than $10^8$. When the condition number of $R_k$ becomes larger than $10^{10}$, the relative residual norm starts to increase sharply. This observation shows that when the condition number of $R_k$
becomes very large, the backward substitution will fail to give an accurate $y_k$. As a result, we would not get an accurate $x_k$, and the convergence of AB-GMRES would deteriorate.

4 Stabilized GMRES method

In this section, we first propose and present a stabilized GMRES method. Then, we explain its regularization effect comparing it with other regularization techniques.

4.1 The stabilized GMRES

In order to overcome the deterioration of convergence of GMRES for inconsistent systems, we propose solving the normal equations

$$ R_k^T R_k y_k = R_k^T t_k $$

instead of $R_k y_k = t_k$ of (15), which we will call the stabilized GMRES. We replace line 8 of Algorithm 1 by Algorithm 2. This makes the system consistent, and stabilizes the process, as will be shown in the following.

One may also consider using the normal equations of $H_{k+1,k}$. However, before breakdown, we use the standard AB-GMRES, which means we do not have to store $H_{k+1,k}$. We only store $R_k$ and update it in each iteration, which is cheaper.

Figure 4 shows the relative residual norm $\|A^T R_k\|_2 / \|A^T R_0\|_2$ versus the number of iterations for the standard AB-GMRES and stabilized AB-GMRES with $B = A^T$ for Maragal_3T. The stabilized method reaches the relative residual norm level of $10^{-11}$ which improves a lot compared to the standard method. The method which we used for solving the normal equations (16) is the standard Cholesky decomposition without pivoting.
This seems paradoxical, since forming the normal equations whose coefficient matrix $R_k^TR_k$ would square the condition number compared to $R_k$, which would make the ill-conditioned problem even worse. Why can the stabilized AB-GMRES give a more accurate solution? We will explain why the stabilized AB-GMRES works in the next subsection.

In spite of the above mentioned merits of stabilization, solving the normal equations in AB-GMRES is expensive. Actually, we only need the stabilized AB-GMRES when $R_k$ becomes ill-conditioned. Thus, we can speed up the process by switching AB-GMRES to stabilized AB-GMRES only when $R_k$ becomes ill-conditioned. The condition number of an incrementaly enlarging triangular matrix can be estimated by techniques in [30]. In this paper, we adopt the switching strategy by monitoring the relative residual norm $\frac{\|A^TR_k\|_2}{\|A^TR_0\|_2}$. Let $\text{ATR}(k) = \frac{\|A^TR_k\|_2}{\|A^TR_0\|_2}$ for the $k$th iteration. When $\text{ATR}(v)\min_{k=1,2,\ldots,v-1}\text{ATR}(k) > 10$, we judge that a jump in relative residual norm has occurred, and we switch AB-GMRES to stabilized AB-GMRES at the $v$th iteration.

4.2 Why the stabilized GMRES method works

Consider solving $R_ky_k = t_k, R_k \in \mathbb{R}^{k \times k}, t_k \in \mathbb{R}^k$ by solving the normal equations (16), which, in theory, squares the condition number and makes the problem become harder to solve numerically. However, in finite precision arithmetic, the condition

![Algorithm 2 Normal equations stabilization approach]

1: Compute the QR decomposition of $H_{k+1,k} = Q_{k+1}R_{k+1,k}$.
2: $R_{k+1,k} = \begin{pmatrix} R_k & 0 \end{pmatrix}^T$, $Q_{k+1}^T\delta e_1 = \begin{pmatrix} t_k \\ \rho_{k+1} \end{pmatrix}$, $\tilde{R}_k = R_k^TR_k$, $\tilde{t}_k = R_k^Tt_k$.
3: Compute the Cholesky decomposition of $R_k = LL^T$.
4: Solve $Lx_k = \tilde{t}_k$ by forward substitution.
5: Solve $L^Ty_k = z_k$ by backward substitution.

Fig. 4 Comparison of the standard AB-GMRES with stabilized AB-GMRES for Maragal_3T
number of the normal equations is not necessarily squared. We will continue to illustrate the phenomenon by using the example in Sect. 3.

We used the MATLAB function `svd` in quadruple precision arithmetic [1] to calculate the singular values. The smallest singular value of $R_{550}$ is $5.39 \times 10^{-15}$, so its square is $2.91 \times 10^{-29}$.

Let $fl(\cdot)$ denote the evaluation of an expression in floating point arithmetic and $fl_d(\cdot)$ and $fl_q(\cdot)$ denote the result in double precision arithmetic and quadruple precision arithmetic, respectively. Figure 5 shows that, numerically, the smallest singular value of $fl_d(R_{s550}^T R_{550})$ is $7.21 \times 10^{-14}$, which is much larger than $2.91 \times 10^{-29}$. Further, the Cholesky factor $L$ of $fl_d(R_{s550}^T R_{550}) = LL^T$ computed in double precision has the smallest singular value $3.50 \times 10^{-7}$, which is also larger than $\sqrt{2.91 \times 10^{-29}} = 5.39 \times 10^{-15}$. Thus, the triangular systems $Lz_k = t_k$ and $L^T y_k = z_k$ are better-conditioned than $R_k y_k = t_k$, which will ensure the stability of the forward and backward substitutions and succeed in obtaining a much more accurate solution with stability compared to the standard approach as shown in Fig. 4.

The left of Fig. 6 compares the singular values $\sigma_k(fl_d(R_{s610}^T R_{610}))$ and $\sigma_k(R_{610})^2$, $k = 1, 2, \ldots, 550$. The first to the 549th singular values of $fl_d(R_{s550}^T R_{550})$ and the corresponding $\sigma(R_{550})^2$ are almost the same, while the last one is different.

What will happen when $R_k$ contains a cluster of small singular values?

The upper triangular matrix $R_{610}$ contains a cluster of small singular values. The right of Fig. 6 compares the singular values $\sigma_k(fl_d(R_{610}^T R_{610}))$ and $\sigma_k(R_{610})^2$. The larger singular values are the same as the ‘exact’ values, while the smaller singular values become larger than the ‘exact’ ones.

Experiment results show that finite precision arithmetic has the effect of shifting the tiny singular values upwards and reduce the condition number of $R^T R$. Besides the fact that the possibly inconsistent system (15) is replaced by the consistent system (16), that is the reason why the normal equations (16) help to make the problem easier to solve.

Next, we computed $R_{s550}^T R_{550}$ in quadruple precision arithmetic and observed that the smallest singular values of $R_{s550}^T R_{550}$ coincided with the squared singular values $\sigma_k(R_{550})^2$ (blue circle symbol) in the left of Fig. 6, unlike in double precision.

![Fig. 5](image_url) Singular values $\sigma_k(fl_d(R_{s550}^T R_{550})), k = 1, 2, \ldots, 550$ in quadruple precision arithmetic

\[ \sigma_k(fl_d(R_{s550}^T R_{550})), k = 1, 2, \ldots, 550 \text{ in quadruple precision arithmetic} \]
computation. Since the maximum of the elements of \( \|f_d(R_{550}^T R_{550}) - f_d(R_{610}^T R_{610}) \|_1 \) is approximately \( 8.16 \times 10^{-12} \), double precision arithmetic contains error of the order of \( 10^{-12} \). Thus, double precision arithmetic has an effect of regularizing the matrix \( R_{550}^T R_{550} \), since double precision matrix multiplication is not accurate enough to keep all the information.

### 4.3 Quadruple precision

In order to see the effect of the machine precision on the convergence of AB-GMRES, we compared the stabilized AB-GMRES with the standard AB-GMRES in quadruple precision arithmetic for the problem Maragal\_3T in Fig. 7 in terms of the relative residual norm \( \|A^T r_k\|_2 / \|A^T b\|_2 \) versus the number of iterations. For both methods, the relative residual norm reached a lower level of order \( 10^{-16} \) compared to \( 10^{-12} \) and \( 10^{-8} \), respectively, for double precision arithmetic in Fig. 4. The curves of the relative residual norm became smoother compared to double precision. As seen

---

Fig. 6  Singular values \( \sigma_k(f_d(R_{550}^T R_{550})) \), \( \sigma_k(R_{550}) \), \( \sigma_k(f_d(R_{610}^T R_{610})) \), and \( \sigma_k(R_{610}) \) in quadruple precision arithmetic

Fig. 7  Effect of the stabilized method in quadruple precision arithmetic for Maragal\_3T
in Fig. 7, the relative residual norm of the standard AB-GMRES jumped to $10^{-1}$ after reaching $10^{-16}$, whereas the relative residual norm of the stabilized GMRES stayed around $10^{-16}$.

### 4.4 Rounding error analysis of the stabilized GMRES method

In order to understand the stability of the proposed method, we perform rounding error analysis. Let $u$ be the unit roundoff [14], which is about $1.11 \times 10^{-16}$ for the IEEE 754 binary 64 (double) in our experiments. The analysis shows that if the condition number of $R \in \mathbb{R}^{n \times n}$ is $\frac{1}{\alpha(n \sqrt{u})}$, then the condition number of $R^TR$ can be reduced from $\frac{1}{\alpha(n^2u)}$ in exact arithmetic to $O\left(\frac{1}{n^2u}\right)$ in finite precision arithmetic.

Let $fl(x)$ denote the floating point number corresponding to $x$. Let $A = (a_{ij}) \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, and $|A| = (|a_{ij}|)$.

Then, we have (cf. [14])

$$|fl(AB) - AB| \leq \gamma_n |A||B|, \quad \text{where} \quad \gamma_n := \frac{nu}{1 - nu}. \quad (17)$$

Hence,

$$|fl(R^TR) - R^TR| \leq \gamma_n |R^T||R|. \quad (18)$$

Let

$$E = (\epsilon_{ij}) := fl(R^TR) - R^TR. \quad (19)$$

Then, we have

$$|E| = (|\epsilon_{ij}|) \leq \gamma_n |R^T||R|. \quad (20)$$

Let $R = U\Sigma V^T$ be the singular value decomposition (SVD) of $R \in \mathbb{R}^{n \times n}$, where $U = [u_1, u_2, \ldots, u_n]$, $V = [v_1, v_2, \ldots, v_n] \in \mathbb{R}^{n \times n}$ are orthogonal matrices, and $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$.

Then, note the following.

**Lemma 6** ($|R^T||R|$)$_{ij} \leq \|R\|_2^2 = \sigma_1^2$.

**Proof** Let $R = [r_1, r_2, \ldots, r_n]$. Then,

$$(|R^T||R|)_{ij} = |r_i|^T r_j = (|r_i|, |r_j|) \leq \|r_i\|_2 \|r_j\|_2 \leq \max_{1 \leq i \leq n} \|r_i\|_2^2$$

$$= \max_{i=1, \ldots, n} \|Re_i\|_2^2 \leq \max_{\|x\|_2 = 1} \|Rx\|_2^2 \leq \|R\|_2^2 = \sigma_1^2.$$ 

Here, $e_i$ is the $i$th column of the identity matrix. \hfill \square
Hence, we have
\[
|\epsilon_{ij}| \leq \gamma_n(|R^T||R|)_{ij} \leq \gamma_n \sigma_1^2. \quad (21)
\]
Let \( A, B, C \) be \( n \times n \) Hermitian matrices, and \( A = B + C \). Denote the eigenvalues of \( A, B \) and \( C \) by \( \lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A) \), \( \lambda_1(B) \geq \lambda_2(B) \geq \cdots \geq \lambda_n(B) \), and \( \lambda_1(C) \geq \lambda_2(C) \geq \cdots \geq \lambda_n(C) \), respectively. Then, the following Weyl’s inequality (See e.g. [15, 16])
\[
\lambda_k(B) + \lambda_n(C) \leq \lambda_k(A) \leq \lambda_k(B) + \lambda_1(C), \quad k = 1, 2, \ldots, n, \quad (22)
\]
holds.
Hence, we have
\[
|\lambda_k(A) - \lambda_k(B)| \leq \|C\|_2 = \|A - B\|_2, \quad k = 1, 2, \ldots, n. \quad (23)
\]
Letting \( A = \tilde{f}(R^T R), B = R^T R, C = E \), we have
\[
|\lambda_k(\tilde{f}(R^T R)) - \lambda_k(R^T R)| \leq \|E\|_2, \quad k = 1, 2, \ldots, n. \quad (24)
\]
Let \( \tilde{f}(R^T R) = \tilde{V}_k \tilde{\Sigma}_k \tilde{V}_k^T \) be the SVD of \( \tilde{f}(R^T R) \), where \( \tilde{\Sigma} = \text{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_n) \), \( \tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \cdots \geq \tilde{\sigma}_n \geq 0 \), which gives
\[
|\tilde{\sigma}_k^2 - \sigma_k^2| \leq \|E\|_2, \quad k = 1, 2, \ldots, n. \quad (25)
\]
Note
\[
|\tilde{\sigma}_k^2 - \sigma_k^2| \leq \|E\|_2 \leq \|E\|_F
\]
\[
= \sqrt{\sum_{i=1}^n |\epsilon_{ij}|^2} \leq \sqrt{\sum_{i,j=1}^n (\gamma_n \sigma_1^2)^2}
\]
\[
= \sqrt{n^2 (\gamma_n \sigma_1^2)^2} = n \gamma_n \sigma_1^2, \quad k = 1, 2, \ldots, n. \quad (26)
\]
Hence,
\[
|\tilde{\sigma}_k^2 - \sigma_k^2| \leq n \gamma_n \sigma_1^2, \quad k = 1, 2, \ldots, n, \quad (27)
\]
i.e.
\[
\sigma_k^2 - n \gamma_n \sigma_1^2 \leq \tilde{\sigma}_k^2 \leq \sigma_k^2 + n \gamma_n \sigma_1^2, \quad k = 1, 2, \ldots, n, \quad (28)
\]
or
\[
\tilde{\sigma}_k^2 = \sigma_k^2 + t_k n \gamma_n \sigma_1^2, \quad -1 \leq t_k \leq 1, \quad k = 1, 2, \ldots, n. \quad (29)
\]
Recall \( \gamma_n = \frac{nu}{1-nu} u \approx 1.11 \times 10^{-16} \). If \( nu \ll 1 \iff n \ll \frac{1}{u} \approx 9.01 \times 10^{15} \) for double precision arithmetic. Then,
\[
\frac{1}{1 - nu} \approx 1 + nu \implies \gamma_n = \frac{nu}{1 - nu} \approx nu(1 + nu) \approx nu.
\] (30)

Hence,
\[
\tilde{\sigma}_k^2 \approx \sigma_k^2 + t_k n^2 u \sigma_1^2, \quad -1 \leq t_k \leq 1, \quad k = 1, 2, \ldots, n.
\] (31)

Then,
\[
\tilde{\sigma}_1^2 \approx \sigma_1^2(1 + t_1 n^2 u), \quad -1 \leq t_1 \leq 1.
\] (32)

We define the following Landau’s symbols:

\[
f(x) = O(g(x)) \text{ as } x \to a \text{ denotes that } \frac{f(x)}{g(x)} \text{ is bounded as } x \to a,
\] (33)

and

\[
f(x) = o(g(x)) \text{ as } x \to a \text{ denotes that } \lim_{x \to a} \frac{f(x)}{g(x)} = 0.
\] (34)

In the following, for instance, \(o(n \sqrt{u})\) is defined by letting \(x = n \sqrt{u}, a = 0\).

Assume \(n \sqrt{u} \ll 1 (\iff n \ll \frac{1}{\sqrt{u}} \approx 9.49 \times 10^7)\). Then, since, \(\frac{1}{\sqrt{u}} \ll \frac{1}{u}\), we have \(n \ll \frac{1}{\sqrt{u}} \ll \frac{1}{u}\). Thus, \(nu \ll 1\).

Assume

\[
k = \kappa(R) = \frac{\sigma_1}{\sigma_n} = \frac{1}{o(n \sqrt{u})}
\] (35)

\(\iff \kappa^2 = \frac{1}{o(n^2 u)} \iff \frac{1}{\kappa^2} = o(n^2 u)\)

holds. Then,

\[
\tilde{\sigma}_n^2 \approx \sigma_n^2 + t_n n^2 u \sigma_1^2 = \sigma_1^2 \left( \frac{\sigma_n^2}{\sigma_1^2} + t_n n^2 u \right)
\] (36)

\[
= \sigma_1^2 \left( \frac{1}{\kappa^2} + t_n n^2 u \right) \approx \sigma_1^2 \left[ o(n^2 u) + t_n n^2 u \right].
\]

Assume \(|t_n| > o(1)\). (Note that if we assume \(t_n\) is randomly distributed in the interval \([-1, 1]\), then, generically, \(o(1) < |t_n|\) holds.) Then, since \(-1 \leq t_n \leq 1, o(1) < |t_n| = O(1)\) holds. Hence, \(\tilde{\sigma}_n^2 \approx \sigma_1^2 t_n n^2 u\). Since \(\tilde{\sigma}_n^2 \geq 0\), we have \(t_n > 0\) and \(\frac{1}{t_n} = O(1)\).

Hence, from (32), (36) and \(\frac{1}{t_n} = O(1)\), we have
\[ \tilde{\kappa}^2 = \kappa(\text{fl}(R^T R)) = \frac{\tilde{\sigma}^2}{\sigma^2_n} \approx \frac{1}{t_n n^2 u} \approx O\left(\frac{1}{n^2 u}\right). \]  

(37)

In summary, we have the following theorem.

**Theorem 7** Let \( u \) be the unit roundoff, and \( R \in \mathbb{R}^{n \times n} \). If \( n\sqrt{u} \ll 1 \) and \( \kappa(R) = \frac{\sigma_1(R)}{\sigma_n(R)} = \frac{1}{o(n\sqrt{u})} \), then, generically, \( \sigma_n(\text{fl}(R^T R)) \approx \sigma_1(R)^2 t_n n^2 u \), where \( o(1) < t_n = O(1) \), and \( \kappa(\text{fl}(R^T R)) = O\left(\frac{1}{n^2 u}\right) \) hold.

**Remark 1** For IEEE double \( n \approx 1.11 \times 10^{-16} \), \( n\sqrt{u} \ll 1 \iff n \ll 9.49 \times 10^7 \).

**Remark 2** \( \kappa(R^T R) = \frac{1}{o(n^2 u)} \).

Then, numerical experiments suggest that if \( LL^T \) is the Cholesky decomposition of \( \text{fl}(R^T R) \) computed in finite precision, then \( \kappa(L) = O\left(\frac{1}{n\sqrt{u}}\right) \), even when \( \kappa(R) = \frac{1}{o(n\sqrt{u})} \).

Thus, forming the normal equations and applying Cholesky decomposition can lead to a more stable computation for extremely ill-conditioned systems of equations, and hence explains why the stabilized GMRES method works without choosing the value of a regularization parameter such as in TSVD or Tikhonov regularization, which will be mentioned in §5.1.1 and 5.1.2, respectively.

Let us compare estimates with numerical results for the Maragal-3T matrix in Fig. 6.

For \( R_{550} \), \( n = 550 \), \( \sigma_1(R_{550}) \approx 1.90 \times 10^2 \), \( \sigma_{550}(R_{550}) \approx 5.39 \times 10^{-15} \). Hence,

\[ \kappa(R_{550}) = \frac{\sigma_1(R_{550})}{\sigma_{550}(R_{550})} \approx 3.53 \times 10^{16} = \frac{1}{o(n\sqrt{u})} \gg \frac{1}{n\sqrt{u}} \approx 1.73 \times 10^5. \]

Thus, \( \tilde{\sigma}^2_n \approx \sigma^2_n t_n n^2 u \approx 1.21 \times 10^{-6} \), where \( o(1) < t_n = O(1) \), and \( \kappa(\text{fl}(R^T R)) \approx O\left(\frac{1}{n^2 u}\right) \approx 2.98 \times 10^{10} \), whereas in Fig. 6, \( \tilde{\sigma}^2_n \approx 7.21 \times 10^{-14} \), and \( \kappa(\text{fl}(R^T R)) \approx 5.00 \times 10^{17} \).

For \( R_{610} \), \( n = 610 \), \( \sigma_1(R_{610}) \approx 2.13 \times 10^2 \), \( \sigma_{610}(R_{610}) \approx 2.91 \times 10^{-15} \). Hence,

\[ \kappa(R_{610}) = \frac{\sigma_1(R_{610})}{\sigma_{610}(R_{610})} \approx 7.32 \times 10^{16} = \frac{1}{o(n\sqrt{u})} \gg \frac{1}{n\sqrt{u}} \approx 1.56 \times 10^5. \]

Thus, \( \tilde{\sigma}^2_n \approx \sigma^2_n t_n n^2 u \approx 1.87 \times 10^{-6} \), where \( o(1) < t_n = O(1) \), and \( \kappa(\text{fl}(R^T R)) \approx O\left(\frac{1}{n^2 u}\right) \approx 2.42 \times 10^{10} \), whereas in Fig. 6, \( \tilde{\sigma}^2_n \approx 1.62 \times 10^{-14} \), and \( \kappa(\text{fl}(R^T R)) \approx 2.77 \times 10^{18} \).

We summarize the results in Table 2. We think there are two reasons for the overestimation of \( \tilde{\sigma}^2_n \). One comes from the inequality \( \|E\|_2 \leq \|E\|_F \) in (26). The other is that \( t_n > o(1) \), but \( t_n \) may be considerably smaller than 1 in (29).
We remark that [31] analyzes the stability of the CholeskyQR2 algorithm using similar techniques. However, they assume $\kappa(R) = O\left(\frac{1}{\sqrt{\alpha}}\right)$, whereas we assume $\kappa(R) = o\left(\frac{1}{\sqrt{\alpha}}\right)$.

### 4.5 Two advantages of forming the normal equations

When $R$ is singular, $R^{-1}$ does not exist, and

$$Ry = t \quad (38)$$

does not have a solution when $t \notin \mathcal{R}(R)$.

If we reformulate (38) as a least squares problem

$$\min_y \|t - Ry\|_2, \quad (39)$$

then (39) has a solution even when $t \notin \mathcal{R}(R)$. For instance, the minimum-norm solution of (39) is given by $y = R^t t$, where $R^t$ is the pseudo-inverse of $R$.

Note that (39) is equivalent to the normal equations

$$R^T Ry = R^t t. \quad (40)$$

(40) is consistent, i.e. $R^T t \in \mathcal{R}(R^T) = \mathcal{R}(R^T R)$, and has a solution.

Now consider the case when $R$ is nearly singular (severely ill-conditioned), that is $\kappa(R) = o\left(\frac{1}{\sqrt{\alpha}}\right)$. Then, solving (38) by backward substitution fails to give an accurate solution as shown in Fig. 3.

On the other hand, we may expect that we may obtain a numerical solution of the least squares problem (39), for instance by approximating $y = R^t t$ [29].

In fact, since (39) is equivalent to the normal equations, as we have seen in Theorem 7, $\kappa(\text{fl}(R^T R)) = O\left(\frac{1}{\sqrt{\alpha}}\right)$ holds while $\kappa(R^T R) = \frac{1}{o(\alpha^{1/4})}$, which gives a numerical advantage.

Thus, we may say that forming the normal equations (40) has two advantages over the system of equations (38). One is that, it makes the system consistent and guarantees the existence of a solution, which opens the possibility of a numerical solution by some
kind of approximation. The other advantage is that normal equations become numerically better conditioned than in exact arithmetic.

4.6 When the stabilized GMRES method works

The stabilized GMRES does not always stabilize the solution of the upper triangular system. A counter example is when \( R_k \) is a Läuchli matrix \[13\], implying that \( R_k/R_{k+3} \) computed in finite precision becomes singular. Indeed, when GMRES is applied to a linear system with an EP (equal preojecton) matrix \( A_3 \), that is \( N(A_3) = N(A_3^T) \) such as

\[
A_3 x = \begin{pmatrix}
\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{6u}}{6} \\
\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} + \frac{\sqrt{6u}}{6} & \frac{\sqrt{6u}}{6} \\
0 & \frac{\sqrt{6u}}{3} & \frac{\sqrt{6u}}{3}
\end{pmatrix} x = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \tag{41}
\]

where \( A_3 \) has the null space \( N(A_3) = \text{span}\{(1, -1, 1)^T\} \), and \( u \) is the unit roundoff, the resulting \( R_k \) is a Läuchli matrix.

Apply GMRES with \( x_0 = 0 \) to (41). Let \( R_k \in \mathbb{R}^{k \times k} \) be the upper triangular matrix obtained at the \( k \)th iteration of GMRES. In the second iteration, after applying the Givens rotation to \( H_3^{1,2} \), we obtain the following:

\[
R_2 = \begin{pmatrix} 1 & 1 \\ 0 & \sqrt{u} \end{pmatrix}, \quad R_2^T R_2 = \begin{pmatrix} 1 & 1 & 1+u \\ 1 & 1 \end{pmatrix} \approx \begin{pmatrix} 1 & 1 \end{pmatrix}. \tag{42}
\]

Thus, there is a risk that the stabilized GMRES will give a numerically singular matrix \( R_2^T R_2 \) in finite precision arithmetic for nonsingular \( R_2 \). We will analyze this phenomenon.

Note that the following theorem holds from Theorem 8.10 of \[14\], where \( |b| = (|b_1|, |b_2|, \ldots, |b_n|)^T \) for \( b = (b_1, b_2, \ldots, b_n)^T \in \mathbb{R}^n \).

**Theorem 8** Let \( T = (t_{ij}) \in \mathbb{R}^{n \times n} \) be a triangular matrix and \( b \in \mathbb{R}^n \). Then, the computed solution \( \hat{x} \) obtained from substitution applied to \( Tx = b \) satisfies

\[
\hat{x} = x + O(n^2 u M(T)^{-1}) |b|. \tag{43}
\]

Here, \( M(T) = (m_{ij}) \) is the comparison matrix such that

\[
m_{ij} = \begin{cases} |t_{ij}|, & i = j, \\ -|t_{ij}|, & i \neq j. \end{cases} \tag{44}
\]

Further, we define the following. Let

\[
\mathcal{O}(x) = \begin{pmatrix} O(x) \\ O(x) \\ \vdots \\ O(x) \end{pmatrix} \in \mathbb{R}^n, \quad \mathcal{O}(x) = [\mathcal{O}(x), \mathcal{O}(x), \ldots, \mathcal{O}(x)] \in \mathbb{R}^{n \times n}. \tag{45}
\]
We assume that the basic arithmetic operations \( \text{op} = +, -, \star, / \) satisfy \( \text{fl}(x \text{ op } y) = (x \text{ op } y)(1 + O(u)) \) as in \([14]\).

Let \( x, y \in \mathbb{R}^n, A \in \mathbb{R}^{n \times n} \). Then,

\[
\text{fl}(x^T y) = x^T y + O(nu)|x|^T|y| = x^T y + O(nu), \\
\text{fl}(Ax) = Ax + O(nu)|A||x| = Ax + O(nu).
\]

Let \( C \in \mathbb{R}^{n \times n} \) and \( \|C\|_2 = O(1) \). We say \( C \in \mathbb{R}^{n \times n} \) is numerically nonsingular if the statement

\[
\text{fl}(Cx) = O(u) \quad \Rightarrow \quad x = O(u)
\]  \( \text{(46)} \)

holds. Note that this definition of numerical nonsingularity agrees with that of numerical rank \([2]\) due to the following.

Let the SVD of \( C = U \Sigma V^T \), where \( U, V \) are orthogonal matrices and \( \Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n) \). We assume \( \|C\|_2 = \sigma_1 = O(1) \). If the numerical rank of \( C \) is \( r < n \), there is a singular value \( \sigma_i = O(u) \), \( r + 1 \leq i \leq n \). Then, \( Cx = U \Sigma V^T x = O(u) \) admits \( x' = V^T x = (x'_1, x'_2, \ldots, x'_n)^T \) such that \( x'_i = O(1) \), and hence \( x = O(1) \). Thus, \( C \) is numerically singular. Then, the following theorem holds.

**Theorem 9** Let \( R_k = (r_{ij}) \in \mathbb{R}^{k \times k} \) be an upper-triangular matrix and

\[
R_{k+1} = \begin{pmatrix} R_k & d \\ 0^T & r_{k+1,k+1} \end{pmatrix} \in \mathbb{R}^{(k+1) \times (k+1)}.
\]  \( \text{(47)} \)

Assume that \( R_k \) is nonsingular and numerically nonsingular, \( R_k = O(1), R_k^{-1} = O(1), M(R_k)^{-1} = O(1), d = O(1) \), and \( O(k) = O(k^2) = O(1) \). Then, the following holds:

\[
\text{fl}(R_{k+1}^T R_{k+1}) \text{ is numerically nonsingular } \iff \text{ fl}(r_{k+1,k+1}^2) > \text{ fl}(d^T d) O(u).
\]

**Proof** See Appendix 3.

Theorem 9 gives the necessary and sufficient condition so that the stabilized GMRES works at the \((k + 1)\)st iteration, i.e. \( R_{k+1}^T R_{k+1} \) is numerically nonsingular.

The difficulty in solving \( R_{k+1} y_i = t_i \) by backward substitution is not necessarily because the diagonals of \( R_i \) are tiny. The reason is that \( R_i \) has tiny singular values. However, the exceptional example (42) exists where the stabilized AB-GMRES does not work. The condition \( \text{fl}(r_{k+1,k+1}^2) > \text{ fl}(d^T d) O(u) \) in Theorem 9 excludes such exceptions.

Figure 8 shows \( r_{k+1,k+1}^2 \) and \( d^T d \) together with the relative residual norm \( \|A^T r_k\|_2 / \|A^T b\|_2 \) of AB-GMRES and stabilized AB-GMRES for Maragal_3T. The figure shows that up to 613 iterations, the conditions in Theorem 9 are satisfied, and \( R_{k+1}^T R_{k+1} \) is numerically nonsingular, so that the stabilized AB-GMRES works.

In fact, for
of \((42), \sigma_1(R_2) \approx \sqrt{2}, \sigma_2(R_2) \approx \sqrt{\frac{u}{2}}, \) so that \(\kappa(R_2) \approx \frac{2}{\sqrt{u}} \approx O\left(\frac{1}{n\sqrt{u}}\right) \ll o\left(\frac{1}{n\sqrt{u}}\right), \) so the condition \((35)\) is not satisfied, and the stabilized GMRES is not guaranteed to work in this case.

5 Comparisons with other methods

We show the numerical performance of the proposed stabilized AB-GMRES method on test matrices, compared with previous methods. All programs for iterative methods were coded according to the algorithms in \([7, 11, 21, 25]\). Each method was terminated at the iteration step which gives the minimum relative residual norm within \(m\) iterations, where \(m\) is the number of the rows of the matrix. No restarts were used for GMRES. Experiments were done for rank-deficient underdetermined matrices whose information is given in Table 1. Here, we have deleted the zero rows and columns of the test matrices beforehand. The elements of \(b\) were randomly generated using the MATLAB function \texttt{rand}. Therefore, generically \(b \notin \mathcal{R}(A)\) and the problem is inconsistent. Each experiment was done 10 times for the same right-hand side \(b\) and the average of the CPU times are shown. The symbol \(-\) denotes that \(\|A^T r_k\|_2/\|A^T b\|_2\) did not reach \(10^{-8}\) within \(m\) iterations. The symbol \((\ast)\) denotes that we used the MATLAB function \texttt{chol} instead of Cholesky decomposition without pivoting for solving the normal equations \((16)\) to save CPU time, except for Havard500, for which Cholesky decomposition without pivoting did not converge. The symbol \((\&)\) denotes the case where even using the MATLAB function \texttt{chol} for solving equation \((16)\) failed to converge. Then, we used the MATLAB function backslash for solving the normal equations \((16)\).
5.1 Underdetermined inconsistent least squares problems

5.1.1 Comparison with Truncated SVD method

Motivated by the stabilized AB-GMRES, we also applied the truncated singular value decomposition (TSVD) stabilization method and compared it with the stabilized AB-GMRES. The method modifies $R_k$ by truncating singular values smaller than $\mu$. More specifically, let $R_k = U\Sigma V^T$ be the SVD of $R_k$, where the columns of $U = [u_1, u_2, \ldots, u_k]$ and $V = [v_1, v_2, \ldots, v_k]$ are the left and right singular vectors, respectively, and the diagonal entries of $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k)$ are the singular values of $R_k$ in descending order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$. Then, the TSVD approximates $R_k \simeq \sum_{j=1}^{k} \frac{1}{\sigma_j}u_jv_j^T$ with $j$ such that $\sigma_{j+1} \leq \mu \sigma_1 \leq \sigma_j$ and $y_k = R_k^{-1}t_k \simeq \sum_{j=1}^{j} \frac{1}{\sigma_j}v_ju_j^Tt_j, j \leq k$.

For the problem Maragal_3T, when $\mu = 10^{-13}, 10^{-12}, \ldots, 10^{-4}$, the method converges but when $\mu$ is smaller than $10^{-13}$ or larger than $10^{-4}$, it does not converge as shown in Fig. 9. Numerical experiments showed that $\mu = \sqrt{u} \simeq 10^{-8}$, where $u$ is the unit roundoff (about $10^{-16}$ in double precision arithmetic), gave the best result among $\mu = 10^{-1}, 10^{-2}, \ldots, 10^{-16}$ in terms of the relative residual norm. The convergence behaviour of the TSVD stabilization method with $\mu = 10^{-8}$ is similar to the stabilized AB-GMRES method as shown in Fig. 10, which suggests that eliminating tiny singular values of $R_k$ which are less than $10^{-8}$ is effective for solving problem (1). However, the TSVD method requires computing the truncated singular value decomposition of $R_k$, and requires choosing the value of the threshold parameter $\mu$, whereas the stabilized AB-GMRES does not require either of them.

5.1.2 Comparison with Tikhonov regularization method

Another approach to stabilize AB-GMRES would be to apply Tikhonov regularization. There are two methods to implement it. The first method is to solve the following square system:

![Fig. 9](image-url)
using the Cholesky decomposition. The second method is to solve the regularized least squares problem

\[(R_k^T R_k + \lambda I)y_k = R_k^T t_k, \quad \lambda \geq 0\]  \hspace{1cm} (48)

using the QR decomposition. These two methods are equivalent mathematically. However, they are not equivalent numerically. The behavior of the first method is similar to the stabilized AB-GMRES.

Table 3 shows that AB-GMRES combined with the first method converges better when \(\lambda = 10^{-16}\) than when \(\lambda = 10^{-14}\) for the problem Maragal\_3T. This method can be used to shift upwards the small singular values, but is less accurate compared to the stabilized AB-GMRES.

Table 3 also shows that the second method is even more accurate compared with the stabilized AB-GMRES method. There is no need to form the normal equations, so that less information is lost due to rounding error. However, one needs to choose an appropriate value for the regularization parameter \(\lambda\). Figure 11 shows the relative residual norm \(\|A^T r_k\|_2/\|A^T r_0\|_2\) for AB-GMRES with Tikhonov regularization using (49) versus the number of iterations for different values of \(\lambda\) for Maragal\_3T. According to Fig. 11, \(\lambda = 10^{-16}\) was optimal among \(10^{-12}, 10^{-14}, 10^{-16}\), and \(10^{-18}\).

We here note the following.

**Theorem 10** Let \(\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k\) be the the singular values of \(R_k\). Then, the singular values of

\[R'_k = \left( \frac{R_k}{\sqrt{\lambda} I} \right)\]  \hspace{1cm} (50)
are given by $\sqrt{\sigma_1^2 + \lambda} \geq \sqrt{\sigma_2^2 + \lambda} \geq \cdots \geq \sqrt{\sigma_k^2 + \lambda}$.

**Proof** See Appendix 4.

Then, let

$$
\kappa \equiv \kappa_2(R_k) = \frac{\sigma_1}{\sigma_k}, \quad \kappa' \equiv \kappa_2(R'_k) = \frac{\sigma_1^2 + \lambda}{\sigma_1^2 + \lambda} = 1 + \frac{\sigma_1^2(1 - 1/\kappa^2)}{\sigma_1^2 + \lambda}. \quad (51)
$$

Since $\kappa \geq 1$, $d\kappa'/d\lambda \leq 0$ for $\lambda \geq 0$ and $\kappa'(\lambda = 0) = \kappa$, $\kappa'(\lambda = +\infty) = 1$. Note also that

**Fig. 11** Relative residual norm for AB-GMRES with Tikhonov regularization using (49) versus number of iterations for different values of the regularization parameter $\lambda$ for Maragal_3T.
\[ \lambda = \frac{\sigma_1^2 \left[ 1 + (\kappa' / \kappa)^2 \right]}{\kappa'^2 - 1}. \]

(52)

Therefore, for instance, if \( \kappa \gg 1 \) and we want \( \kappa' = \sqrt{\kappa} \),
\[ \lambda = \frac{\sigma_1^2 \left( 1 + 1 / \kappa \right)}{\kappa - 1} \approx \frac{\sigma_1^2}{\kappa}. \]

(53)

For example, if \( \kappa = 10^{16} \) and we want \( \kappa' = 10^8 \), we should choose \( \lambda \approx \sigma_1^2 \times 10^{-16} \). For Maragal_3T, the largest singular value \( \sigma_1 \) is about 12.64, so that we can estimate a reasonable value of \( \lambda \approx 1.60 \times 10^{-14} \). However, this estimation assumes \( \kappa' = \sqrt{\kappa} \), and needs an extra cost for computing \( \sigma_1 \). See [3] for other estimation techniques for the regularization parameter.

### 5.1.3 Comparison with the range restricted GMRES

We compared the proposed stabilized AB-GMRES with the range restricted AB-GMRES (RR-AB-GMRES) [21], where the Krylov subspace for the RR-AB-GMRES with \( B = A^\top \) is \( K_k(AA^\top, AA^\top r_0) \), and the standard AB-GMRES with \( B = A^\top \).

Table 4 gives the number of iterations and the smallest relative residual norm for the RR-AB-GMRES, the standard and stabilized AB-GMRES for the Maragal matrices. The table shows that the stabilized AB-GMRES is more accurate than the standard AB-GMRES. Table 4 also shows that the stabilized AB-GMRES is generally more accurate than the RR-AB-GMRES. The stabilized AB-GMRES took more iterations to attain the same order of the smallest residual norm than the RR-AB-GMRES.

### 5.2 Inconsistent systems with severely ill-conditioned range-symmetric coefficient matrices

Next, we test the stabilized AB-GMRES on least squares problems \( \min_{x \in \mathbb{R}^n} \| b - Ax \|_2 \) by GMRES, where \( A \in \mathbb{R}^{n \times n} \) are severely ill-conditioned range-symmetric (square) matrices given in Table 5.

#### Table 4 Comparison of the attainable smallest relative residual norm \( \| A^\top r_k \|_2 / \| A^\top r_0 \|_2 \)

| Matrix       | Maragal_3T | Maragal_4T | Maragal_5T | Maragal_6T | Maragal_7T |
|--------------|------------|------------|------------|------------|------------|
| Iter.        | 531        | 465        | 1110       | 2440       | 1864       |
| Standard AB-GMRES | 1.05x10^{-8} | 2.09x10^{-7} | 5.35x10^{-6} | 8.26x10^{-6} | 4.53x10^{-6} |
| Iter.        | 552        | (&) 598    | (+) 1224   | (+) 3000   | (+) 2475   |
| Stabilized AB-GMRES | 4.86x10^{-12} | 5.59x10^{-8} | 2.54x10^{-6} | 4.56x10^{-6} | 2.78x10^{-7} |
| Iter.        | 553        | 565        | 1223       | 2374       | 2474       |
| RR-AB-GMRES  | 2.57x10^{-11} | 5.59x10^{-8} | 3.62x10^{-6} | 1.63x10^{-5} | 2.78x10^{-7} |
These matrices are all numerically singular. We generated the right-hand side $b$ by the MATLAB function `rand`, so that the systems are generically inconsistent. We compared the stabilized AB-GMRES with the standard AB-GMRES and RR-AB-GMRES. Table 6 gives the smallest relative residual norm and the corresponding number of iterations. Table 7 gives the CPU times in seconds required to obtain relative residual norm $\|A^T r_k\|_2/\|A^T r_0\|_2 < 10^{-8}$. The switching strategy which was introduced in Sect. 4.1 was used for the stabilized AB-GMRES when measuring CPU times. The number of iterations when switching occurred is in brackets.

For Harvard500 and bw42, AB-GMRES could only converge to the level of $10^{-9}$ regarding the relative residual norm, while the stabilized AB-GMRES converged to the level of $10^{-14}$. The stabilized AB-GMRES was robust in the sense that it could continue to compute even when the upper triangular matrix $R_k$ became seriously ill-conditioned, and the relative residual norm did not increase sharply towards the end, but just stagnated at a low level, just like for consistent problems.

Thus, our stabilization method also makes AB-GMRES stable for highly ill-conditioned inconsistent systems with square coefficient matrices.

The coefficient matrix $A$ of bw42 is singular and satisfies $\mathcal{N}(A) = \mathcal{N}(A^T)$. The problem comes from a finite-difference discretization of a PDE with periodic boundary condition (Experiment 4.2 in Brown and Walker[4] with the original $b$). Since the matrix is range symmetric, the GMRES, RR-GMRES, and stabilized GMRES can be directly applied to $Ax = b$ (See [4] Theorem 2.4, [10] Theorem 2.7, and [5] Theorem 3.2.) as shown in Table 8. The stabilized GMRES gave a relative residual norm $1.94 \times 10^{-11}$ for bw42 at the 219th iteration. The proposed method can be considered as a way of making the GMRES stable for highly ill-conditioned inconsistent problems.

Figure 12 shows comparison of GMRES with stabilized GMRES for a symmetric matrix (, which is range symmetric), freeFlyingRobot\_7 which contains a cluster of tiny singular values which gradually decrease to zero. The stabilized GMRES converged to $3.65 \times 10^{-10}$ at 3,452 iterations, better than GMRES. But the relative residual increased after the 3,452 iterations. Hence, we adopted a reorthogonalization strategy which performs the modified Gram-Schmidt orthogonalization process once more. We replaced line 4-6 of Algorithm 1 (GMRES

---

**Table 5 Information of the singular square matrices**

| Matrix            | Size  | density[\%] | rank  | $\kappa_2(A)$ | application                      |
|-------------------|-------|--------------|-------|---------------|----------------------------------|
| Harvard500        | 500   | 1.05         | 170   | $1.30 \times 10^2$ | Web connectivity                 |
| netz4504          | 1961  | 0.13         | 1342  | $3.41 \times 10^4$ | 2D/3D finite element problem     |
| TS                | 2142  | 0.99         | 2140  | $3.52 \times 10^3$ | Counter example problem          |
| grid2\_dual       | 3136  | 0.12         | 3134  | $8.58 \times 10^3$ | 2D/3D finite element problem     |
| uk                | 4828  | 0.06         | 4814  | $6.62 \times 10^3$ | Undirected graph                 |
| bw42              | 10000 | 0.05         | 9999  | $2.03 \times 10^4$ | Partial differential equation[4] |
| msc01050          | 1050  | 2.38         | 1049  | $1.31 \times 10^8$ | 2D/3D structural problem         |
| freeFlyingRobot\_7| 3918  | 0.20         | 3881  | $1.68 \times 10^{12}$ | Optimal control problem          |
Table 6  Comparison of the attainable smallest relative residual norm $\|A^Tr_i\|_2/\|A^Tr_0\|_2$ for inconsistent square linear systems

| Matrix               | Harvard500 | netz4504 | TS     | grid2_dual | uk     | bw42 |
|----------------------|------------|----------|--------|------------|--------|------|
| Iter.                | 104        | 144      | 1487   | 3134       | 4620   | 742  |
| Standard AB-GMRES    | $9.38 \times 10^{-9}$ | $4.51 \times 10^{-10}$ | $1.56 \times 10^{-9}$ | $5.98 \times 10^{-10}$ | $1.35 \times 10^{-9}$ | $8.06 \times 10^{-8}$ |
| Iter.                | (+) 134    | (&) 201  | 1613   | (+) 3135   | (+) 4739 | (&) 788 |
| Stabilized AB-GMRES  | $8.46 \times 10^{-14}$ | $1.51 \times 10^{-14}$ | $2.51 \times 10^{-9}$ | $5.53 \times 10^{-10}$ | $6.57 \times 10^{-10}$ | $1.66 \times 10^{-7}$ |
| Iter.                | 135        | 200      | 1652   | 3134       | 4706   | 1163 |
| RR-AB-GMRES          | $7.78 \times 10^{-14}$ | $3.36 \times 10^{-14}$ | $4.56 \times 10^{-9}$ | $6.52 \times 10^{-8}$ | $8.33 \times 10^{-8}$ | $1.56 \times 10^{-5}$ |
version) by Algorithm 3 to reorthogonalize GMRES and the stabilized GMRES. As in Fig. 13, after reorthogonalization, the stabilized GMRES became more stabilized and converged to a relative residual of $6.45 \times 10^{-11}$ at 3,701 iterations.

**Table 7** Comparison of the CPU time (seconds) to obtain relative residual norm $\|A^T r_k \|_2 / \|A^T r_0 \|_2 < 10^{-8}$ for inconsistent square linear systems

| matrix         | Harvard500 | netz4504 | TS   | grid2_dual | uk   | bw42 |
|----------------|------------|----------|------|------------|------|------|
| Iter.          | 104        | 134      | 1411 | 3134       | 4583 | –    |
| Standard AB-GMRES | 4.72×10^{-2} | 1.87×10^{-1} | 2.14×10 | 2.16×10^{2} | 6.93×10^{2} | – |
| Iter.          | 104        | 134      | 1531 (182) | 3134 | 4679 (4199) | – |
| Stabilized AB-GMRES | 4.78×10^{-2} | 1.89×10^{-1} | 8.19×10^{2} | 2.21×10^{2} | 1.93×10^{3} | – |
| Iter.          | 114        | 153      | 1530 | – | – | – |
| RR-AB-GMRES    | 6.42×10^{-2} | 2.62×10^{-1} | 2.68×10 | – | – | – |

**Table 8** Attainable smallest relative residual norm $\|A^T r_k \|_2 / \|A^T r_0 \|_2$ for range symmetric matrices

| Matrix         | bw42     | msc01050 | FreeFlyingRobot_7 |
|----------------|----------|----------|-------------------|
| Iter.          | 147      | 560      | 1084              |
| Standard GMRES | 8.08×10^{-9} | 4.98×10^{-8} | 8.86×10^{-8} |
| Iter.          | 219      | 668      | 3414              |
| Stabilized GMRES | 2.11×10^{-11} | 4.62×10^{-9} | 3.24×10^{-10} |
| Iter.          | 220      | 564      | 3183              |
| RR-GMRES       | 3.13×10^{-11} | 2.62×10^{-6} | 1.40×10^{-9} |

**Fig. 12** Comparison of GMRES with stabilized GMRES for freeFlyingRobot_7
Algorithm 3 reorthogonalized modified Gram-Schmidt

1: for $i = 1, 2$ do
2:     for $j = 1, 2, \ldots, k$ do
3:         $h_{j,k} = w_k^T v_j$, $w_k = w_k - h_{j,k} v_j$
4:     end for
5: end for

6 Concluding remarks

We proposed a stabilized AB-GMRES method for ill-conditioned underdetermined and inconsistent least squares problems. It shifts upwards the tiny singular values of the upper triangular matrix appearing in AB-GMRES, making the process more stable, giving better convergence, and more accurate solutions compared to AB-GMRES. We have also given a theoretical analysis to explain why the proposed method works. The method is also effective for making GMRES stable for range-symmetric inconsistent least squares problems with severely ill-conditioned square coefficient matrices.

Appendix 1 Proof of statement in section 2.3

Lemma 11 Assume $\mathcal{N}(\hat{A}) \cap \mathcal{R}(\hat{A}) = \{0\}$, and grade$(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = k$. Then, $K_{k+1}(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = \hat{A}K_k(\hat{A}, b|_{\mathcal{R}(\hat{A})})$ holds.

Proof Note that

Fig. 13 Comparison of reorthogonalized GMRES with reorthogonalized stabilized GMRES for freeFlyingRobot_7
\[ \hat{A} \mathcal{K}_k(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = \text{span}\{\hat{A}b|_{\mathcal{R}(\hat{A})}, \hat{A}^2b|_{\mathcal{R}(\hat{A})}, \ldots, \hat{A}^k b|_{\mathcal{R}(\hat{A})}\} \]
\[ \subseteq \text{span}\{b|_{\mathcal{R}(\hat{A})}, \hat{A}b|_{\mathcal{R}(\hat{A})}, \ldots, \hat{A}^k b|_{\mathcal{R}(\hat{A})}\} = \mathcal{K}_{k+1}(\hat{A}, b|_{\mathcal{R}(\hat{A})}). \]

\[ \text{grade}(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = k \] implies that

\[ \mathcal{K}_{k+1}(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = \mathcal{K}_k(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = \text{span}\{b|_{\mathcal{R}(\hat{A})}, \hat{A}b|_{\mathcal{R}(\hat{A})}, \ldots, \hat{A}^{k-1} b|_{\mathcal{R}(\hat{A})}\}. \]

Hence,

\[ \hat{A}^k b|_{\mathcal{R}(\hat{A})} = c_0 b|_{\mathcal{R}(\hat{A})} + c_1 \hat{A}b|_{\mathcal{R}(\hat{A})} + \cdots + c_{k-1} \hat{A}^{k-1} b|_{\mathcal{R}(\hat{A})}, \quad c_i \in \mathbb{R}, \, i = 0, 1, 2, \ldots, k - 1. \]

If \( c_0 = 0 \),

\[ \hat{A}^k b|_{\mathcal{R}(\hat{A})} = c_1 \hat{A}b|_{\mathcal{R}(\hat{A})} + c_2 \hat{A}^2 b|_{\mathcal{R}(\hat{A})} + \cdots + c_{k-1} \hat{A}^{k-1} b|_{\mathcal{R}(\hat{A})}. \]

Hence,

\[ c_1 \hat{A}b|_{\mathcal{R}(\hat{A})} + c_2 \hat{A}^2 b|_{\mathcal{R}(\hat{A})} + \cdots + c_{k-1} \hat{A}^{k-1} b|_{\mathcal{R}(\hat{A})} = \hat{A}(c_1 b|_{\mathcal{R}(\hat{A})} + \cdots + c_{k-1} \hat{A}^{k-2} b|_{\mathcal{R}(\hat{A})}) = 0. \]

Hence,

\[ c_1 b|_{\mathcal{R}(\hat{A})} + c_2 \hat{A}^2 b|_{\mathcal{R}(\hat{A})} + \cdots + c_{k-1} \hat{A}^{k-2} b|_{\mathcal{R}(\hat{A})} \in \mathcal{N}(\hat{A}) \cap \mathcal{R}(\hat{A}) = \{0\}. \]

which implies

\[ \hat{A}^{k-1} b|_{\mathcal{R}(\hat{A})} = c_1 b|_{\mathcal{R}(\hat{A})} + c_2 \hat{A}b|_{\mathcal{R}(\hat{A})} + \cdots + c_{k-1} \hat{A}^{k-2} b|_{\mathcal{R}(\hat{A})}. \]

Thus,

\[ \mathcal{K}_k(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = \mathcal{K}_{k-1}(\hat{A}, b|_{\mathcal{R}(\hat{A})}), \]

which contradicts with \( \text{grade}(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = k \). Hence, \( c_0 \neq 0 \), and

\[ b|_{\mathcal{R}(\hat{A})} = d_1 \hat{A}b|_{\mathcal{R}(\hat{A})} + d_2 \hat{A}^2 b|_{\mathcal{R}(\hat{A})} + \cdots + d_{k-1} \hat{A}^{k-1} b|_{\mathcal{R}(\hat{A})} + d_k \hat{A}^k b|_{\mathcal{R}(\hat{A})}. \]

Hence,

\[ \mathcal{K}_{k+1}(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = \text{span}\{b|_{\mathcal{R}(\hat{A})}, \hat{A}b|_{\mathcal{R}(\hat{A})}, \ldots, \hat{A}^k b|_{\mathcal{R}(\hat{A})}\} \]
\[ \subseteq \text{span}\{\hat{A}b|_{\mathcal{R}(\hat{A})}, \hat{A}^2 b|_{\mathcal{R}(\hat{A})}, \ldots, \hat{A}^k b|_{\mathcal{R}(\hat{A})}\} = \hat{A} \mathcal{K}_k(\hat{A}, b|_{\mathcal{R}(\hat{A})}). \]

Thus,

\[ \mathcal{K}_{k+1}(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = \hat{A} \mathcal{K}_k(\hat{A}, b|_{\mathcal{R}(\hat{A})}). \]

**Corollary 12** Assume \( \mathcal{N}(\hat{A}) = \mathcal{N}(\hat{A}^T) \), and \( \text{grade}(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = k \). Then, \( \mathcal{K}_{k+1}(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = \hat{A} \mathcal{K}_k(\hat{A}, b|_{\mathcal{R}(\hat{A})}) \) holds.
Proof \( \mathcal{N}(\hat{A}) = \mathcal{N}(\hat{A}^T) \) implies that

\[
\mathcal{N}(\hat{A}) \cap \mathcal{R}(\hat{A}) = \mathcal{N}(\hat{A}^T) \cap \mathcal{R}(\hat{A}) = \mathcal{R}(\hat{A})^\perp \cap \mathcal{R}(\hat{A}) = \{0\}.
\]

Hence, from Lemma 11, Corollary 12 holds.

Appendix 2 Proof of statement in section 2.3

Lemma 13 Assume \( \mathcal{N}(\hat{A}) \cap \mathcal{R}(\hat{A}) = \{0\} \), \( \text{grade}(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = k \), and \( b \notin \mathcal{R}(\hat{A}) \). Then, \( \dim(\mathcal{K}_{k+1}(\hat{A}, b)) = k + 1 \) holds.

Proof Let \( c_0, c_1, \ldots, c_k \in \mathbb{R} \) satisfy

\[
c_0b + c_1\hat{A}b + \cdots + c_k\hat{A}^kb = 0.
\]

Since \( \mathcal{N}(\hat{A}) \cap \mathcal{R}(\hat{A}) = \{0\} \),

\[
b = b|_{\mathcal{R}(\hat{A})} \oplus b|_{\mathcal{N}(\hat{A})},
\]

where \( b|_{\mathcal{N}(\hat{A})} \) denotes the orthogonal projection of \( b \) onto \( \mathcal{N}(\hat{A}) \). Hence,

\[
c_0b|_{\mathcal{N}(\hat{A})} + c_0b|_{\mathcal{R}(\hat{A})} + c_1\hat{A}b|_{\mathcal{R}(\hat{A})} + \cdots + c_k\hat{A}^kb|_{\mathcal{R}(\hat{A})} = 0.
\]

If \( c_0 \neq 0 \)

\[
b|_{\mathcal{N}(\hat{A})} = -\frac{c_1}{c_0}\hat{A}b|_{\mathcal{R}(\hat{A})} - \cdots - \frac{c_k}{c_0}\hat{A}^kb|_{\mathcal{R}(\hat{A})} \in \mathcal{R}(\hat{A}).
\]

Hence,

\[
b|_{\mathcal{N}(\hat{A})} \in \mathcal{N}(\hat{A}) \cap \mathcal{R}(\hat{A}) = \{0\}.
\]

Thus, \( b|_{\mathcal{N}(\hat{A})} = 0 \), which contradicts \( b \notin \mathcal{R}(\hat{A}) \). Hence, we have \( c_0 = 0 \), and

\[
c_1\hat{A}b + c_2\hat{A}^2b + \cdots + c_k\hat{A}^kb = c_1\hat{A}b|_{\mathcal{R}(\hat{A})} + c_2\hat{A}^2b|_{\mathcal{R}(\hat{A})} + \cdots + c_k\hat{A}^kb|_{\mathcal{R}(\hat{A})} = 0.
\]

But, since

\[
\dim(\text{span}(\hat{A}b|_{\mathcal{R}(\hat{A})}, \hat{A}^2b|_{\mathcal{R}(\hat{A})}, \cdots, \hat{A}^kb|_{\mathcal{R}(\hat{A})})) = \dim(\hat{A}\text{span}(b|_{\mathcal{R}(\hat{A})}, \hat{A}b|_{\mathcal{R}(\hat{A})}, \cdots, \hat{A}^{k-1}b|_{\mathcal{R}(\hat{A})})) = \dim(\hat{A}\mathcal{K}_k(\hat{A}, b|_{\mathcal{R}(\hat{A})})) = k
\]

holds from Lemma 11, we have \( c_1 = c_2 = \cdots = c_k = 0 \), which implies \( \dim(\mathcal{K}_{k+1}(\hat{A}, b)) = k + 1 \).

Corollary 14 Assume \( \mathcal{N}(\hat{A}) = \mathcal{N}(\hat{A}^T) \), \( \text{grade}(\hat{A}, b|_{\mathcal{R}(\hat{A})}) = k \), and \( b \notin \mathcal{R}(\hat{A}) \).

Then, \( \dim(\mathcal{K}_{k+1}(\hat{A}, b)) = k + 1 \) holds.
Proof \( \mathcal{N}(\hat{A}) = \mathcal{N}(\hat{A}^\top) \) implies \( \mathcal{N}(\hat{A}) \cap \mathcal{R}(\hat{A}) = \{0\} \). Hence, the corollary follows from Lemma 13.

Appendix 3 Proof of Theorem 9 in section 4.6

Theorem 9 Let \( R_k = (r_{pq}) \in \mathbb{R}^{k \times k} \) be an upper-triangular matrix and

\[
R_{k+1} = \begin{pmatrix} R_k & d \\ 0 & r_{k+1,k+1} \end{pmatrix} \in \mathbb{R}^{(k+1) \times (k+1)}.
\]

(54)

Assume \( R_k \) is numerically nonsingular, and \( R_k = \mathcal{O}(1), \ R_k^{-1} = \mathcal{O}(1), \ M(R_k)^{-1} = \mathcal{O}(1), \ d = \mathcal{O}(1) \) and \( O(k) = O(k^2) = O(1) \). Then, the following holds:

\[ \text{fl}(R_{k+1}^TR_{k+1}) \text{ is numerically nonsingular } \iff \text{fl}(r_{k+1,k+1}^2) > \text{fl}(d^Td)O(u). \]

Proof Note that

\[
R_{k+1}^TR_{k+1} = \begin{pmatrix} R_k & 0 \\ d^T & r_{k+1,k+1} \end{pmatrix} \begin{pmatrix} R_k & d \\ 0 & r_{k+1,k+1} \end{pmatrix} = \begin{pmatrix} R_k^TR_k & R_k^Td \\ d^TR_k & d^Td + r_{k+1,k+1}^2 \end{pmatrix}.
\]

Proof of \((\Rightarrow)\)

Assume \( \text{fl}(r_{k+1,k+1}^2) \leq \text{fl}(d^Td)O(u) \). Then, since

\[
\text{fl}(d^Td) = d^Td + O(ku)d^Td = (1 + O(ku))d^Td,
\]

\[
\text{fl}(d^Td + r_{k+1,k+1}^2) = (d^Td + r_{k+1,k+1}^2)(1 + O(ku)) = d^Td(1 + O(ku)),
\]

\[
R_k = \mathcal{O}(1), \quad \text{and} \quad d = \mathcal{O}(1),
\]

we have

\[
\text{fl}(R_{k+1}^TR_{k+1}) = \begin{pmatrix} R_k^TR_k + O(ku)|R_k|^T|R_k| & R_k^Td + O(ku)|R_k|^T|d| \\ d^TR_k + O(ku)|d|^T|R_k| & d^Td + O(ku)d^Td \end{pmatrix}
\]

(55)

\[
= \begin{pmatrix} R_k^T \\ d^T \end{pmatrix} \begin{pmatrix} R_k & d \end{pmatrix} + O(ku).
\]

Note

\[
\begin{pmatrix} R_k & d \end{pmatrix} \begin{pmatrix} -R_k^{-1}d \\ 1 \end{pmatrix} = -R_kR_k^{-1}d + d = 0,
\]

since \( R_k \) is nonsingular.

Hence,

\[
\text{fl} \left( \begin{pmatrix} R_k & d \end{pmatrix} \begin{pmatrix} -R_k^{-1}d \\ 1 \end{pmatrix} \right) = \text{fl} \{ R_k \text{fl}(-R_k^{-1}d) + d \} = [\text{fl} \{ R_k \text{fl}(-R_k^{-1}d) \} + d] \{ 1 + O(u) \}.
\]

Note here that
\[
\mathcal{f}(R_k \mathcal{f}(-R_k^{-1} d)) = R_k \mathcal{f}(-R_k^{-1} d) + O(ku)|R_k||R_k^{-1}d|,
\]
and
\[
\mathcal{f}(-R_k^{-1} d) = -R_k^{-1}d + O(k^2 u)M(R_k)^{-1}|d|
\]  
(56)

from Theorem 8. Hence,
\[
\mathcal{f}(RT_k R_{k+1} \begin{pmatrix} -R_k^{-1}d \\ 1 \end{pmatrix}) = O(k^2 u)R_k M(R_k)^{-1}|d| + O(ku)|R_k||R_k^{-1}d| = O(k^2 u),
\]
since \(R_k^{-1} = O(1)\) and \(M(R_k)^{-1} = O(1)\).

Then,
\[
\mathcal{f}(R_{k+1}^T R_{k+1} \begin{pmatrix} -R_k^{-1}d \\ 1 \end{pmatrix}) = \mathcal{f}(\begin{pmatrix} R_k^T \\ d^T \end{pmatrix}(R_k d) + O(ku)) \begin{pmatrix} -R_k^{-1}d + O(k^2 u)M(R_k)^{-1}|d| \\ 1 \end{pmatrix} = O(k^2 u) = O(u),
\]
since (55), (56), and \(O(k^2) = O(1)\). Since \(\begin{pmatrix} -R_k^{-1}d \\ 1 \end{pmatrix} = O(1)\), \(R_{k+1}^T R_{k+1}\) is numerically singular. By contraposition, \((\Rightarrow)\) holds.

Proof of \((\Leftarrow)\)

Assume \(R_{k+1}^T R_{k+1}\) is not numerically nonsingular. Then, there exists a vector
\[
\begin{pmatrix} z \\ w \end{pmatrix} \in \mathbb{R}^{k+1} \text{ such that } \left\| \begin{pmatrix} z \\ w \end{pmatrix} \right\| > O(u), \text{ and }
\]
\[
\mathcal{f}(R_{k+1}^T R_{k+1} \begin{pmatrix} z \\ w \end{pmatrix}) = R_{k+1}^T \left( R_{k+1} \begin{pmatrix} z \\ w \end{pmatrix} + |R_{k+1}| \left\| \begin{pmatrix} z \\ w \end{pmatrix} \right\| O((k+1)u) \right) +
\]
\[
\left| R_{k+1}^T \right| R_{k+1} \begin{pmatrix} z \\ w \end{pmatrix} + |R_{k+1}| \left\| \begin{pmatrix} z \\ w \end{pmatrix} \right\| O((k+1)u) O((k+1)u) = O(u)
\]
assuming \(O(k+1) = O(1)\).

Hence,
\[
\mathcal{f}(R_{k+1}^T R_{k+1} \begin{pmatrix} z \\ w \end{pmatrix}) = \begin{pmatrix} R_k^T R_k \\ d^T R_k \end{pmatrix} \begin{pmatrix} R_k^T \end{pmatrix} d + r_{k+1,k+1}^2 \begin{pmatrix} z \\ w \end{pmatrix} + O(u) = O(u).
\]

Thus,
\[
R_k^T R_k z + wR_k^T d = O(u),
\]  
(57)
\[
d^T R_k z + (d^T d + r_{k+1,k+1}^2)w = O(u).
\]  
(58)
can be expressed as $R_k^T(R_s z + wd) = O(u)$. From Lemma 15, $R_k^T$ is numerically nonsingular, so that
\begin{equation}
R_s z + wd = O(u).
\end{equation}
Hence, from (58), $d^T R_s z + w( d^T d + r_{k+1,k+1}^2 ) = d^T (R_s z + wd) + wr_{k+1,k+1}^2 = O(u)$. Thus, $wr_{k+1,k+1}^2 = O(u)$. If $w = O(u)$, $R_s z = O(u)$ from (59). Since $R_k$ is numerically nonsingular, $z = O(u)$, which contradicts with the assumption.

Hence, $|w| > O(u)$, so that $r_{k+1,k+1}^2 = O(u)$, which gives
\[ \text{fl}(r_{k+1,k+1}^2) = O(u) \leq \text{fl}(d^T d) O(u). \]

**Lemma 15** Let $n = O(1)$. If $A \in \mathbb{R}^{n \times n}$ is numerically nonsingular, and $A^{-1} = O(1)$, then $A^T$ is numerically nonsingular.

**Proof** If
\[ \text{fl}(A^T x) = A^T x + O(nu)|A^T||x| = O(nu), \]
then
\[ \text{fl}(x^T A) = x^T A + O(nu) = O(nu). \]
Thus,
\[ \text{fl}(x^T A y) = \text{fl}(x^T A y) + O(nu)|\text{fl}(x^T A)||y| = O(nu) \]
holds for all $y = O(1)$.

For arbitrary $z = O(1) \in \mathbb{R}^n$, let
\[ y = A^{-1} z = O(1). \]
Then,
\[ \text{fl}(Ay) = Ay + O(nu)|A||y| = z + O(nu)|A||y|. \]
Hence,
\[ z = \text{fl}(Ay) + O(nu)|A||y| = \text{fl}(Ay) + O(nu). \]
Thus, we have
\[ \text{fl}(x^T z) = x^T z + O(nu)|x|^T|z| = \text{fl}(x^T Ay) + O(nu) = O(nu) \]
for arbitrary $z = O(1) \in \mathbb{R}^n$. Hence, $x = O(u)$, so that $A^T$ is numerically nonsingular.
Proof of Theorem 10 in section 5.1.2

Proof Let the singular value decomposition of $R_k$ be given by $R_k = U \Sigma V^T \in \mathbb{R}^{k \times k}$, where $U$, $V$ are orthogonal matrices and $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k)$. Let $I_k \in \mathbb{R}^{k \times k}$ be the identity matrix. Then, we have $R'_k = \left( \frac{R_k}{\sqrt{\lambda I_k}} \right) = U' \Sigma' V^T$, where $U' = \begin{pmatrix} U & 0 \end{pmatrix}$ and $\Sigma' = \begin{pmatrix} \Sigma \sqrt{\lambda I_k} \end{pmatrix}$. Since $\Sigma' \Sigma' = \Sigma^2 + \lambda I_k = \text{diag}(\sigma_1^2 + \lambda, \sigma_2^2 + \lambda, \ldots, \sigma_k^2 + \lambda)$, the singular values of $\frac{R_k}{\sqrt{\lambda I_k}}$ are $\sqrt{\sigma_1^2 + \lambda} \geq \sqrt{\sigma_2^2 + \lambda} \geq \cdots \geq \sqrt{\sigma_k^2 + \lambda}$.

Acknowledgements We would like to thank Dr. Hiroshi Murakami and Professor Lothar Reichel for valuable comments. Ken Hayami was supported by JSPS KAKENHI Grant Number JP15K04768. Keiichi Morikuni was supported by JSPS KAKENHI Grant Number JP16K17639 and JP20K14356 and Hattori Hokokai Foundation. Jun-Feng Yin was supported by the National Natural Science Foundation of China (No. 11971354).

References

1. ADVANPIX LLC.: Multiprecision Computing Toolbox for MATLAB. https://www.advanpix.com/. Version 4.4.5.12711
2. Björck, Å.: Numerical Methods for Least Squares Problems. SIAM. Philadelphia, PA (1996)
3. Brezinski, C., Rodriguez, G., Seatzu, S.: Error estimates for the regularization of least squares problems. Numer. Algorithms 51(1), 61–76 (2009)
4. Brown, P., Walker, H.: GMRES on (nearly) singular systems. SIAM J. Matrix Anal. Appl. 18(1), 37–51 (1997)
5. Calvetti, D., Lewis, B., Reichel, L.: GMRES-type methods for inconsistent systems. Linear Algebra Appl. 316(1–3), 157–169 (2000)
6. Davis, T., Hu, Y.: The University of Florida sparse matrix collection. ACM Trans. Math. Software 38(1), 1–25 (2011)
7. Fong, D.C.L., Saunders, M.: LSMR: An iterative algorithm for sparse least-squares problems. SIAM J. Sci. Comput. 33(5), 2950–2971 (2011)
8. Foster, L.: San Jose State University Singular Matrix Database. http://www.math.sjsu.edu/singular/matrice/
9. Hansen, P.: Discrete Inverse Problems: Insight and Algorithms. SIAM. Philadelphia, PA (2010)
10. Hayami, K., Sugihara, M.: A geometric view of Krylov subspace methods on singular systems. Numer. Linear Algebra Appl. 18(3), 449–469 (2011)
11. Hayami, K., Yin, J., Ito, T.: GMRES methods for least squares problems. SIAM J. Matrix Anal. Appl. 31(5), 2400–2430 (2010)
12. Hestenes, M., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Research Nat. Bur. Standards 49(6), 409–436 (1952)
13. Higham, N.J.: The Test Matrix Toolbox for MATLAB (version 3.0). University of Manchester, Manchester (1995)
14. Higham, N.J.: Accuracy and Stability of Numerical Algorithms, 2nd edn. SIAM. Philadelphia, PA (2002)
15. Horn, R.A., Johnson, C.R.: Matrix Analysis, 2nd edn. Cambridge University Press, New York, NY (2012)
16. Iri, M.: General Theory of Linear Algebra. Asakura, (in Japanese) (2009)
17. Meza, J.C., Symes, W.W.: Deflated Krylov subspace methods for nearly singular linear systems. J. Optim. Theory Appl. 72(3), 441–457 (1992)
18. Morikuni, K.: Inner-iteration Preconditioning for Least Squares Problems. Doctoral Thesis, Department of Informatics, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies (2013)

19. Morikuni, K., Hayami, K.: Convergence of inner-iteration GMRES methods for rank-deficient least squares problems. SIAM J. Matrix Anal. Appl. 36(1), 225–250 (2015)

20. Morikuni, K., Rozložník, M.: On GMRES for singular EP and GP systems. SIAM J. Matrix Anal. Appl. 39(2), 1033–1048 (2018)

21. Neuman, A., Reichel, L., Sadok, H.: Algorithms for range restricted iterative methods for linear discrete ill-posed problems. Numer. Algorithms 59(2), 325–331 (2012)

22. Neuman, A., Reichel, L., Sadok, H.: Implementations of range restricted iterative methods for linear discrete ill-posed problems. Linear Algebra Appl. 436(10), 3974–3990 (2012)

23. Paige, C., Rozložník, M., Strakoš, Z.: Modified Gram-Schmidt (mgs), least squares, and backward stability of MGS-GMRES. SIAM J. Matrix Anal. Appl. 28(1), 264–284 (2006)

24. Paige, C.C., Saunders, M.A.: Solution of sparse indefinite systems of linear equations. SIAM J. Numer. Anal. 12(4), 617–629 (1975)

25. Paige, C.C., Saunders, M.A.: LSQR: An algorithm for sparse linear equations and sparse least squares. ACM Trans. Math. Software 8(1), 43–71 (1982)

26. Reichel, L., Ye, Q.: Breakdown-free GMRES for singular systems. SIAM J. Matrix Anal. Appl. 26(4), 1001–1021 (2005)

27. Saad, Y.: Iterative Methods for Sparse Linear Systems, 2nd edn. SIAM. Philadelphia, PA (2003)

28. Saad, Y., Schultz, M.H.: GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. 7(3), 856–869 (1986)

29. Sugihara, K., Hayami, K., Liao, Z.: GMRES using pseudo-inverse for range symmetric singular systems. (in revision)

30. Tebbens, J.D., Tůma, M.: On incremental condition estimators in the 2-norm. SIAM J. Anal. Appl. 35(1), 174–197 (2014)

31. Yamamoto, Y., Nakatsuksa, Y., Yanagisawa, Y., Fukuya, T.: Roundoff error analysis of the CholeskyQR2 algorithm. Electron. Trans. Numer. Anal. 44(01), 306–326 (2015)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Zeyu Liao1 · Ken Hayami2 · Keiichi Morikuni3 · Jun-Feng Yin4

Ken Hayami
hayami@nii.ac.jp

Keiichi Morikuni
morikuni@cs.tsukuba.ac.jp

Jun-Feng Yin
yinjf@tongji.edu.cn

1 Department of Informatics, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies (SOKENDAI), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

2 National Institute of Informatics, and The Graduate University for Advanced Studies (SOKENDAI), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

3 Faculty of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

4 School of Mathematical Science, Tongji University, Siping Road 1239, Yangpu District, Shanghai 200092, China

Springer