



Research article

Estimation of the backward errors for scaled total least squares problems

Zhanshan Yang^{1,*}, Bing Zheng² and Tiexiang Li³

¹ School of Mathematics and Statistic, Qinghai Minzu University, Xining 810007, China

² School of Mathematics and Statistic, Lanzhou University, Lanzhou 730000, China

³ School of Mathematics, Southeast University, Nanjing 210096, China

* **Correspondence:** Email: yangzhanshan123456@126.com.

Abstract: In this paper, we prove that both the extended minimal backward errors and the true minimal backward errors of the scaled total least squares (STLS) problem are identical to those of its core problem. This conclusion also holds for the asymptotic estimates of the extended minimal backward errors. Benefiting from its lower dimensionality, the core problem can effectively reduce the computational cost of backward error evaluation. We further derive practical, low-cost computable backward error estimates for a STLS problem using the Lanczos bidiagonalization process, and we demonstrate that our results can be readily extended to standard least squares and data least squares problems. To improve the efficiency of the Lanczos bidiagonalization-based TLS algorithm, we propose practical stopping criteria for the iterative numerical solution of STLS problems. Numerical experiments fully validate the computational advantages of our proposed approaches.

Keywords: scaled total least squares problem; core problem; Krylov subspace; Lanczos bidiagonalization; backward error

1. Introduction

Consider estimating x from the real linear system

$$Ax \approx b \text{ for } A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, \text{ and } x \in \mathbb{R}^n, \quad (1.1)$$

where we exclude the trivial case for the sake of clarity via a natural assumption,

$$b \notin \mathcal{R}(A), \text{ that is, } A^T b \neq 0. \quad (1.2)$$

For (1.1), an important example of an optimal approximation problem whose optima are invariant under orthogonal transformations, followed by scaled total least squares (STLS), which can be expressed as

follows: for a given $\gamma > 0$,

$$\text{STLS distance} \equiv \sigma_S \equiv \min_{E, f, x} \|E, \gamma f\|_F \text{ s.t. } (A + E)x = b + f. \quad (1.3)$$

Here, the relative sizes of the corrections in A and b are determined by the real parameter $\gamma > 0$. As $\gamma \rightarrow 0$, the STLS solution approaches the least squares (LS) solution, when $\gamma = 1$ (5) coincides with the total least squares (TLS) formulation, and as $\gamma \rightarrow \infty$, it approaches data least squares (DLS), see [1–3] and so on. The STLS framework unifies ordinary least squares (OLS), DLS, and TLS, moreover, the parameter γ bears statistical interpretations, as discussed in [2, 4]. The STLS problem (1.3) admits a unique solution if the following two conditions hold:

$$\text{rank}(A) = n, \text{ and } b \notin \mathcal{U}_{\min},$$

where $\mathcal{U}_{\min}$ denotes the left singular vector subspace of A associated with its smallest singular value $\sigma_{\min}(A)$.

Let (1.1) transform to $\tilde{A}\tilde{x} \equiv (P^T A Q)(Q^T)x \approx P^T b \equiv \tilde{b}$, where $P^{-1} = P^T$, $Q^{-1} = Q^T$, or equivalently,

$$P^T \begin{pmatrix} b & A \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & Q^T \end{pmatrix} \begin{pmatrix} -1 \\ x \end{pmatrix} = \begin{pmatrix} \tilde{b}_1 & \tilde{A}_{11} & 0 \\ 0 & 0 & \tilde{A}_{22} \end{pmatrix} \begin{pmatrix} -1 \\ \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} \approx 0. \quad (1.4)$$

We state that $\tilde{A}_{11}\tilde{x}_1 \approx \tilde{b}_1$ constitutes the core problem associated with $Ax \approx b$ if the pair $(\tilde{b}_1, \tilde{A}_{11})$ is minimally dimensioned (or equivalently, $\tilde{A}_{22}$ is maximally dimensioned) subject to constraint (1.4). Consequently, redundant information (including singular value multiplicities) and irrelevant data are segregated into the submatrix $\tilde{A}_{22}$; see [3, 5].

Equation (1.2) rules out the pathological case where $\tilde{A}_{11}$ in (1.4) has zero columns. The approximation problem (1.1) can then be recast in terms of solving the core subproblem $\tilde{A}_{11}\tilde{x}_1 \approx \tilde{b}_1$ with $\tilde{x}_2 = 0$. Core problems possess favorable properties such as minimal dimensionality. When a scaled total least squares problem is solved by first finding a core problem, the resulting theory is consistent with earlier generalizations but is much simpler and clearer. The above framework may be leveraged for the solution of ill-posed problems. Paige [5] demonstrated that a decomposition of the form (1.4) can be computed directly by selecting orthogonal matrices P and Q to reduce the augmented matrix (b, A) to a real upper bidiagonal matrix; see also [2], for instance. Partitioning $P = (P_1, P_2)$ and $Q = (Q_1, Q_2)$ is performed by terminating the bidiagonalization process at the first zero entry, which yields relation (1.4), where $\tilde{A}_{22}$ remains unprocessed by bidiagonalization. In other words, for any linear system $Ax \approx b$, the orthogonal upper bidiagonalization of (b, A) yields a corresponding core problem.

Backward perturbation analysis serves two main purposes. First, it verifies whether a computed solution is backward stable: If the numerical stability of an algorithm is unknown, yet its relative minimal backward error is of the order of unit round-off, then the solution is backward stable and acceptable. Second, its outcomes facilitate the construction of effective stopping criteria for iterative solutions to large sparse problems; see, e.g., [6–10]. For the latest achievements, please refer to [11, 12]. In this paper, we prove that the extended minimal backward errors and the true minimal backward errors of the STLS problem (1.1) are equal to those associated with its core subproblem. Identical conclusions hold for the asymptotic estimates of the extended minimal backward errors. Consequently, evaluating the backward error of the core subproblem corresponding to the original STLS problem directly yields

the backward error of the STLS problem itself. This approach is valuable, as the core subproblem possesses far smaller dimensions relative to the original STLS formulation. These theoretical derivations are presented in Section 2. In addition, evaluating the backward error of the approximate solution to the STLS problem entails substantial computational overhead for large-scale sparse systems, making direct computation infeasible. A further contribution of this work lies in deriving computationally inexpensive estimates for this backward error, which can be evaluated at each iteration of Lanczos bidiagonalization, as elaborated in Section 3. We further discuss practical backward error estimation techniques in Section 4. The STLS framework unifies OLS, TLS, and DLS. The minimal backward error estimates for STLS derived herein also unify the backward perturbation analysis frameworks for OLS, TLS, and DLS. Accordingly, this paper generalizes and extends the results in [13], where low-cost, readily computable backward error estimates for ordinary least squares were established. Within the iterative Lanczos bidiagonalization procedure, the quantity ζ_k quantifies the accuracy of the rank- k low-rank approximation of A and can be adopted as a stopping criterion: The iteration terminates once $\zeta_k \leq \text{tol}$, where tol denotes a user-specified tolerance. Against this backdrop, an efficient low-cost scheme for computing ζ_k ($k = 1, 2, \dots$) is proposed in Section 5, which carries significant practical value. We first prove that ζ_k is a monotonically decreasing function of k and admits recursive computation. The proposed estimates for practical stopping criteria can be obtained with nearly negligible extra computational cost within the Lanczos bidiagonalization TLS algorithm, albeit at the expense of d supplementary iterations for evaluation. The computational advantages of all proposed results are fully validated via numerical experiments.

Matrices are denoted by uppercase Roman letters such as $A, B, \dots$; vectors by lowercase Roman letters $v, w, \dots$; and scalars by lowercase Greek letters $\alpha, \beta, \dots$. We use $\|\cdot\|$ to denote the 2-norm of any matrix or vector and $\|\cdot\|_F$ for the Frobenius norm of a matrix A . $A^\dagger$ stands for the Moore–Penrose pseudoinverse of A . The symbol e_j represents the j th standard basis column of the identity matrix I .

2. Backward errors of the STLS problems and the corresponding core problems

2.1. Extended minimal backward error

In order to find the closest STLS problem whose solution is actually y , where $y \in R^n$ is a nonzero approximation to the solution vector x of (1.3), we need to solve the following backward error problem of the form

$$\min_{\Delta A, \Delta b} \|(\Delta A, \Delta b)\theta\|_F \quad \text{s.t.} \quad y = \arg \min_x \frac{\|b + \Delta b - (A + \Delta A)x\|^2}{\gamma^{-2} + \|x\|^2}$$

with a scaling parameter $\theta > 0$. From [7], y is the exact STLS solution for the data set $(A + \Delta A, b + \Delta b)$ if and only if $(\Delta A, \Delta b)$ is in the following set:

$$\begin{aligned} \mathcal{S}_{A,b}(\gamma) &\equiv \left\{ (\Delta A, \Delta b) : (A + \Delta A)^T(b + \Delta b - (A + \Delta A)y) \right. \\ &\quad \left. = -\frac{\|b + \Delta b - (A + \Delta A)x\|^2}{\gamma^{-2} + \|x\|^2}y, \frac{\|b + \Delta b - (A + \Delta A)x\|^2}{\gamma^{-2} + \|x\|^2} < \sigma_{\min}^2(A + \Delta A) \right\}. \end{aligned} \quad (2.1)$$

Because (2.1) is difficult to derive a general expression for $(\Delta A, \Delta b) \in \mathcal{S}_{A,b}(\gamma)$, there need to consider a larger set

$$\mathcal{S}_{A,b}^+(\gamma) \equiv \{(\Delta A, \Delta b) : (A + \Delta A)^T(b + \Delta b - (A + \Delta A)y) = -\frac{\|b + \Delta b - (A + \Delta A)x\|^2}{\gamma^{-2} + \|x\|^2}y\}$$

and give the extended minimal backward error problem

$$\eta(y, \gamma, \theta) \equiv \min_{(\Delta A, \Delta b) \in \mathcal{S}_{A,b}^+(\gamma)} \|(\Delta A, \Delta b\theta)\|_F, \quad (2.2)$$

which seeks to minimize the weighted Frobenius norm $\|[\Delta A, \Delta b\theta]\|_F$ over all perturbations $(\Delta A, \Delta b)$ belonging to the relaxed constraint set $\mathcal{S}_{A,b}^+(\gamma)$ by discarding the strict-inequality condition required for the exact STLS solution whose minimizer yields the extended minimal backward error $\eta(y, \gamma, \theta)$ and is solved as shown the following result from Theorem 3.1 of [7].

Lemma 2.1. Suppose $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ is a nonzero approximate solution of the STLS problem, $\gamma > 0$, $\theta > 0$, and suppose that (1.4) holds. Let $r \equiv b - Ay$, and

$$\begin{aligned} N &\equiv \begin{pmatrix} A(I - yy^\dagger) & \omega(I - rr^\dagger) & v(Ay + \gamma^2\|y\|^2b) \end{pmatrix}, \\ M &\equiv M(y, \gamma, \theta) \equiv A(I - yy^\dagger)A^T + \frac{\theta^2}{1 + \theta^2\|y\|^2}rr^T \\ &\quad + v^2(Ay + \gamma^2\|y\|^2b) \times (Ay + \gamma^2\|y\|^2b)^T = NN^T - \omega^2I. \end{aligned}$$

Then, M has at most one negative eigenvalue, and

$$\eta^2(y, \gamma, \theta) = \begin{cases} \omega^2, & \text{if } \lambda_{\min}(M) \geq 0; \\ \omega^2 + \lambda_{\min}(M) = \sigma_{\min}^2(N), & \text{if } \lambda_{\min}(M) < 0. \end{cases} \quad (2.3)$$

Furthermore, $\eta(y, \gamma, \theta)$ is given by the backward perturbations in A and b :

$$\Delta A = \begin{cases} \frac{\theta^2}{1 + \theta^2\|y\|^2}ry^T, & \text{if } \lambda_{\min}(M) \geq 0; \\ \frac{\theta^2}{1 + \theta^2\|y\|^2}ry^T - w_*w_*^T \left(A/\frac{\gamma^2(\theta^2 - \gamma^2)^2}{\theta} + \gamma^4\|y\|^2by^T \right. \\ \quad \left. + \frac{\gamma^4 + 2\gamma^4\theta^2\|y\|^2 + \theta^4}{(1 + \theta^2\|y\|^2)(\theta^2 + \gamma^4\|y\|^2)}ry^T \right), & \text{if } \lambda_{\min}(M) < 0. \end{cases} \quad (2.4)$$

$$\Delta b = \begin{cases} -\frac{1}{1 + \theta^2\|y\|^2}r, & \text{if } \lambda_{\min}(M) \geq 0; \\ -\frac{1}{1 + \theta^2\|y\|^2}r - w_*w_*^T \left(\frac{\gamma^2(1 + \gamma^2\|y\|^2)}{\theta^2 + \gamma^4\|y\|^2}b \right. \\ \quad \left. - \frac{(\gamma^2 + \theta^2)(1 + \gamma^2\|y\|^2)}{(1 + \theta^2\|y\|^2)(\theta^2 + \gamma^4\|y\|^2)}r \right), & \text{if } \lambda_{\min}(M) < 0, \end{cases} \quad (2.5)$$

where $\omega = \theta\|r\|/\sqrt{1 + \theta^2\|y\|^2}$, $v = \theta/\sqrt{\theta^2\|y\|^2 + \gamma^4\|y\|^4}$, and w_* is the unit eigenvector of M corresponding to $\lambda_{\min}(M)$.

For the core subproblem $\tilde{A}_{11}\tilde{x}_1 \approx \tilde{b}_1$, associated with Problem (1.1) (here we only present the case when $Ax \approx b$ is s incompatible; see Theorem 2.2 in [7] for details), we can also derive its extended minimal backward error under the same parameters θ and γ .

Theorem 2.1. Suppose $\tilde{A}_{11} \in R^{(p+1) \times p}$, $\tilde{b}_1 \in R^{p+1}$, $\tilde{y}_1 \in R^p$ is a nonzero approximate solution of the STLS problem, $\gamma > 0$, $\theta > 0$, and suppose that (1.4) holds. Let $\tilde{r}_1 \equiv \tilde{b}_1 - \tilde{A}_{11}\tilde{y}_1$, and

$$\begin{aligned} N_1 &\equiv \begin{pmatrix} \tilde{A}_{11}(I - \tilde{y}_1\tilde{y}_1^\dagger) & \omega_1(I - \tilde{r}_1\tilde{r}_1^\dagger) & \nu_1(\tilde{A}_{11}\tilde{y}_1 + \gamma^2\|\tilde{y}_1\|^2\tilde{b}_1) \end{pmatrix}, \\ M_1 &\equiv M_1(\tilde{y}_1, \gamma, \theta) \equiv \tilde{A}_{11}(I - \tilde{y}_1\tilde{y}_1^\dagger)\tilde{A}_{11}^T + \frac{\theta^2}{1 + \theta^2\|\tilde{y}_1\|^2}\tilde{r}_1\tilde{r}_1^T \\ &\quad + \nu_1^2(\tilde{A}_{11}\tilde{y}_1 + \gamma^2\|\tilde{y}_1\|^2\tilde{b}_1) \times (\tilde{A}_{11}\tilde{y}_1 + \gamma^2\|\tilde{y}_1\|^2\tilde{b}_1)^T = N_1N_1^T - \omega_1^2I. \end{aligned}$$

Then, M_1 has at most one negative eigenvalue, and

$$\eta^2(\tilde{y}_1, \gamma, \theta) = \begin{cases} \omega_1^2, & \text{if } \lambda_{\min}(M_1) \geq 0; \\ \omega_1^2 + \lambda_{\min}(M_1) = \sigma_{\min}^2(N_1), & \text{if } \lambda_{\min}(M_1) < 0. \end{cases}$$

Furthermore, $\eta(\tilde{y}_1, \gamma, \theta)$ is given by the following backward perturbations in $\tilde{A}_{11}$ and $\tilde{b}_1$:

$$\Delta A_{11} = \begin{cases} \frac{\theta^2}{1 + \theta^2\|\tilde{y}_1\|^2}\tilde{r}_1\tilde{y}_1^T, & \text{if } \lambda_{\min}(M_1) \geq 0; \\ \frac{\theta^2}{1 + \theta^2\|\tilde{y}_1\|^2}\tilde{r}_1\tilde{y}_1^T - w_{1*}w_{1*}^T \left(\tilde{A}_{11} / \frac{\gamma^2(\theta^2 - \gamma^2)^2}{\theta} + \gamma^4\|\tilde{y}_1\|^2\tilde{b}_1\tilde{y}_1^T \right. \\ \left. + \frac{\gamma^4 + 2\gamma^4\theta^2\|\tilde{y}_1\|^2 + \theta^4}{(1 + \theta^2\|\tilde{y}_1\|^2)(\theta^2 + \gamma^4\|\tilde{y}_1\|^2)}\tilde{r}_1\tilde{y}_1^T \right), & \text{if } \lambda_{\min}(M_1) < 0. \end{cases} \quad (2.6)$$

$$\Delta b_1 = \begin{cases} -\frac{1}{1 + \theta^2\|\tilde{y}_1\|^2}\tilde{r}_1, & \text{if } \lambda_{\min}(M_1) \geq 0; \\ -\frac{1}{1 + \theta^2\|\tilde{y}_1\|^2}\tilde{r}_1 - w_{1*}w_{1*}^T \left(\frac{\gamma^2(1 + \gamma^2\|\tilde{y}_1\|^2)}{\theta^2 + \gamma^4\|\tilde{y}_1\|^2}\tilde{b}_1 \right. \\ \left. - \frac{(\gamma^2 + \theta^2)(1 + \gamma^2\|\tilde{y}_1\|^2)}{(1 + \theta^2\|\tilde{y}_1\|^2)(\theta^2 + \gamma^4\|\tilde{y}_1\|^2)}\tilde{r}_1 \right), & \text{if } \lambda_{\min}(M_1) < 0, \end{cases} \quad (2.7)$$

where $\omega_1 = \theta\|\tilde{r}_1\| / \sqrt{1 + \theta^2\|\tilde{y}_1\|^2}$, $\nu_1 = \theta / \sqrt{\theta^2\|\tilde{y}_1\|^2 + \gamma^4\|\tilde{y}_1\|^4}$, and w_{1*} is the unit eigenvector of M_1 corresponding to $\lambda_{\min}(M_1)$.

The following theorem presents the relations between the extended minimal backward errors and the corresponding perturbations of the problem (1.1) and those of its core problem.

Theorem 2.2. The extended minimal backward error of the problem (1.3) equals the one of its core problem, and the corresponding perturbations have the following relations:

$$\eta(y, \gamma, \theta) = \eta(\tilde{y}_1, \gamma, \theta)$$

$$\Delta A_{11} = P_1^T \Delta A Q_1, \quad \Delta b_1 = P_1^T \Delta b,$$

where $P = \begin{pmatrix} P_1 & P_2 \end{pmatrix}$, $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$, $P_1 \in R^{m \times (p+1)}$, $Q_1 \in R^{n \times p}$; see the details in (1.4).

Proof. We know $A = P\tilde{A}Q^T$, $b = P\tilde{b}$, $y = Q\tilde{y}$, $r = b - Ay = P\tilde{b} - P\tilde{A}Q^TQ\tilde{y} = P(\tilde{b} - \tilde{A}\tilde{y}) = P\tilde{r}$; then, by (1.4) we have

$$\begin{aligned}
M &= P\tilde{A}Q^T(I - Q\tilde{y}\tilde{y}^\dagger Q^T)Q\tilde{A}^T P^T - \frac{\theta^2}{1 + \theta^2\|\tilde{y}\|^2}P\tilde{r}\tilde{r}^T P^T \\
&\quad + \frac{\theta^2}{\theta^2\|\tilde{y}\|^2 + \gamma^4\|\tilde{y}\|^4}(P\tilde{A}Q^T Q\tilde{y} + \gamma^2\|\tilde{y}\|^2 P\tilde{b}) \times (P\tilde{A}Q^T Q\tilde{y} + \gamma^2\|\tilde{y}\|^2 P\tilde{b})^T \\
&= P(\tilde{A}(I - \tilde{y}\tilde{y}^\dagger)\tilde{A}^T - \frac{\theta^2}{1 + \theta^2\|\tilde{y}\|^2}\tilde{r}\tilde{r}^T \\
&\quad + \frac{\theta^2}{\theta^2\|\tilde{y}\|^2 + \gamma^4\|\tilde{y}\|^4}(\tilde{A}\tilde{y} + \gamma^2\|\tilde{y}\|^2\tilde{b}) \times (\tilde{A}\tilde{y} + \gamma^2\|\tilde{y}\|^2\tilde{b}))P^T \\
&= P\left\{\begin{pmatrix} \tilde{A}_{11} & 0 \\ 0 & \tilde{A}_{22} \end{pmatrix}\left[I - \begin{pmatrix} \tilde{y}_1 \\ 0 \end{pmatrix}\begin{pmatrix} \tilde{y}_1^\dagger & 0 \end{pmatrix}\right]\begin{pmatrix} \tilde{A}_{11}^T & 0 \\ 0 & \tilde{A}_{22}^T \end{pmatrix} - \frac{\theta^2}{1 + \theta^2\|\tilde{y}_1\|^2}\begin{pmatrix} \tilde{r}_1 \\ 0 \end{pmatrix}\begin{pmatrix} \tilde{r}_1^T & 0 \end{pmatrix}\right. \\
&\quad \left. + \frac{\theta^2}{\theta^2\|\tilde{y}_1\|^2 + \gamma^4\|\tilde{y}_1\|^4}\left[\begin{pmatrix} \tilde{A}_{11} & 0 \\ 0 & \tilde{A}_{22} \end{pmatrix}\begin{pmatrix} \tilde{y}_1 \\ 0 \end{pmatrix} + \gamma^2\|\tilde{y}_1\|^2\begin{pmatrix} \tilde{b}_1 \\ 0 \end{pmatrix}\right]\right. \\
&\quad \left.\times\left[\begin{pmatrix} \tilde{A}_{11} & 0 \\ 0 & \tilde{A}_{22} \end{pmatrix}\begin{pmatrix} \tilde{y}_1 \\ 0 \end{pmatrix} + \gamma^2\|\tilde{y}_1\|^2\begin{pmatrix} \tilde{b}_1 \\ 0 \end{pmatrix}\right]\right\}P^T \\
&= P\left\{\begin{pmatrix} \tilde{A}_{11}(I - \tilde{y}_1\tilde{y}_1^\dagger)\tilde{A}_{11}^T & 0 \\ 0 & \tilde{A}_{22}\tilde{A}_{22}^T \end{pmatrix} - \frac{\theta^2}{1 + \theta^2\|\tilde{y}_1\|^2}\begin{pmatrix} \tilde{r}_1\tilde{r}_1^T & 0 \\ 0 & 0 \end{pmatrix}\right. \\
&\quad \left. + \frac{\theta^2}{\theta^2\|\tilde{y}_1\|^2 + \gamma^4\|\tilde{y}_1\|^4}\left[\begin{pmatrix} \tilde{A}_{11}\tilde{y}_1 \\ 0 \end{pmatrix} + \gamma^2\|\tilde{y}_1\|^2\begin{pmatrix} \tilde{b}_1 \\ 0 \end{pmatrix}\right]\right. \\
&\quad \left.\times\left[\begin{pmatrix} \tilde{A}_{11}\tilde{y}_1 \\ 0 \end{pmatrix} + \gamma^2\|\tilde{y}_1\|^2\begin{pmatrix} \tilde{b}_1 \\ 0 \end{pmatrix}\right]\right\}P^T \\
&= P\begin{pmatrix} M_1 & 0 \\ 0 & \tilde{A}_{22}\tilde{A}_{22}^T \end{pmatrix}P^T \quad (*).
\end{aligned}$$

If $\lambda_{\min}(M) \geq 0$, then

$$\eta^2(y, \gamma, \theta) = \frac{\theta^2\|r\|^2}{1 + \theta^2\|y\|^2} = \frac{\theta^2\|\tilde{r}\|^2}{1 + \theta^2\|\tilde{y}\|^2} = \frac{\theta^2\|\tilde{r}_1\|^2}{1 + \theta^2\|\tilde{y}_1\|^2} = \eta^2(\tilde{y}_1, \gamma, \theta).$$

If $\lambda_{\min}(M) < 0$, we know from (*), $\lambda_{\min}(\tilde{A}_{22}\tilde{A}_{22}^T) > 0$, and $\rho(M) = \rho(M_1) \cup \rho(\tilde{A}_{22}\tilde{A}_{22}^T)$, where $\rho(M)$ means the set of the eigenvalues of the matrix M . When $w_{1*} = P_1^T w_*$, there comes $\lambda_{\min}(M_1) = \lambda_{\min}(M) < 0$, and

$$\begin{aligned}
\eta^2(y, \gamma, \theta) &= \frac{\theta^2\|r\|^2}{1 + \theta^2\|y\|^2} + \lambda_{\min}(M) = \frac{\theta^2\|\tilde{r}\|^2}{1 + \theta^2\|\tilde{y}\|^2} + \lambda_{\min}(M) \\
&= \frac{\theta^2\|\tilde{r}_1\|^2}{1 + \theta^2\|\tilde{y}_1\|^2} + \lambda_{\min}(M_1) = \eta^2(\tilde{y}_1, \gamma, \theta).
\end{aligned}$$

We know that $\tilde{A}_{11} = P_1^T A Q_1$, $\tilde{b}_1 = P_1^T b$, $\tilde{y}_1 = Q_1^T y$, $\tilde{r}_1 = P_1^T r$, and $w_{1*} = P_1^T w_*$. By substituting the five formulas above into Eqs (2.6) and (2.7) and combining them with Eqs (2.4) and (2.5), we can obtain

$$\Delta A_{11} = P_1^T \Delta A Q_1, \quad \Delta b_1 = P_1^T \Delta b,$$

where $P = \begin{pmatrix} P_1 & P_2 \end{pmatrix}$, $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$, $P_1 \in R^{m \times (p+1)}$, $Q_1 \in R^{n \times p}$. $\square$

Theorem 3.2 in [7] and Theorem 2.2 imply that the true minimal backward error of Problem (1.1) equals that of its core problem for some $\varepsilon > 0$ and all $y, \tilde{y}_1$ satisfying $\|y - x^*\| \leq \varepsilon$, $\|\tilde{y}_1 - x^*\| \leq \varepsilon$.

Theorem 2.3. *Let x^* be the true STLS solution of (1.1), and there exist $\varepsilon > 0$ such that for all y and $\tilde{y}_1$ satisfying $\|y - x^*\| \leq \varepsilon$ and $\|\tilde{y}_1 - x^*\| \leq \varepsilon$, $\eta(y, \gamma, \theta)$ and $\eta(\tilde{y}_1, \gamma, \theta)$ are the true minimal backward errors of the problem (1.1) and its core problem, respectively. At this time, there is also $\eta(y, \gamma, \theta) = \eta(\tilde{y}_1, \gamma, \theta)$.*

Proof. The proof is obvious, and we omit it here. $\square$

2.2. Asymptotic estimate

Computing $\eta(y, \gamma, \theta)$ exactly is expensive, so Chang [7] gives an asymptotic estimate for it with an explicit expression as follows:

$$\varsigma(y, \gamma, \theta) = \|EE^\dagger f\| = \|E(E^T E)^{-1} E^T f\|, \quad (2.8)$$

where

$$E \equiv \begin{pmatrix} A + \xi_0^{-1} \xi_1 r y^T \\ \xi_0^{-2} \|r\| \|y\| (I - y y^\dagger) \\ \theta^{-1} \xi_0^{-2} \|r\| I \end{pmatrix},$$

$$f \equiv \begin{pmatrix} \frac{\theta}{\sqrt{1+\theta^2} \|y\|^2} r \\ 0 \\ \frac{\|r\|(\theta^2 - \gamma^2)}{(1+\gamma^2 \|y\|^2) \sqrt{1+\theta^2} \|y\|^2} y \end{pmatrix},$$

$$\xi_0 \equiv \sqrt{\|y\|^2 + \theta^{-2}}, \quad \xi_1 \equiv \frac{\|y\|^2 - \gamma^{-2} + 2\theta^{-2}}{\alpha \xi_0},$$

and

$$\alpha \equiv \|y\|^2 + \gamma^{-2}.$$

Naturally, we also expect the asymptotic estimate for $\eta(y, \gamma, \theta)$ of Problem (1.1) to be identical to that for $\eta(y_1, \gamma, \theta)$ of the corresponding core problem. The result is as follows:

Theorem 2.4. *The asymptotic estimate for $\eta(y, \gamma, \theta)$ of the problem (1.3) equals the asymptotic estimate for $\eta(\tilde{y}_1, \gamma, \theta)$ of the corresponding core problem, that is,*

$$\varsigma(y, \gamma, \theta) = \varsigma(\tilde{y}_1, \gamma, \theta).$$

Proof. According to (1.4), we have

$$E \equiv \begin{pmatrix} A + \xi_0^{-1} \xi_1 r y^T \\ \xi_0^{-2} \|r\| \|y\| (I - y y^\dagger) \\ \theta^{-1} \xi_0^{-2} \|r\| I \end{pmatrix} = \begin{pmatrix} P \tilde{A} Q^T + \xi_0^{-1} \xi_1 P \tilde{r} \tilde{y}_1^T Q^T \\ \xi_0^{-2} \|\tilde{r}\| \|\tilde{y}_1\| (I - Q \tilde{y}_1 \tilde{y}_1^T Q^T) \\ \theta^{-1} \xi_0^{-2} \|\tilde{r}\| I \end{pmatrix}$$

$$= \begin{pmatrix} P & & \\ & Q & \\ & & Q \end{pmatrix} \begin{pmatrix} \tilde{A}_{11} + \xi_0^{-1} \xi_1 \tilde{r}_1 \tilde{y}_1^T & 0 \\ 0 & \tilde{A}_{22} \\ \xi_0^{-2} \|\tilde{r}_1\| \|\tilde{y}_1\| (I - \tilde{y}_1 \tilde{y}_1^T) & 0 \\ 0 & \xi_0^{-2} \|\tilde{r}_1\| \|\tilde{y}_1\| I \\ \theta^{-1} \xi_0^{-2} \|\tilde{r}_1\| I & 0 \\ 0 & \theta^{-1} \xi_0^{-2} \|\tilde{r}_1\| I \end{pmatrix} Q^T,$$

$$E^T E = Q \begin{pmatrix} E_1^T E_1 & 0 \\ 0 & E_2^T E_2 \end{pmatrix} Q^T,$$

where

$$E_1^T E_1 = (\tilde{A}_{11} + \xi_0^{-1} \xi_1 \tilde{r}_1 \tilde{y}_1^T)^T (\tilde{A}_{11} + \xi_0^{-1} \xi_1 \tilde{r}_1 \tilde{y}_1^T) + (\xi_0^{-2} \|\tilde{r}_1\| \|\tilde{y}_1\|)^2 (I - \tilde{y}_1 \tilde{y}_1^\dagger) + (\theta^{-1} \xi_0^{-2} \|\tilde{r}_1\|)^2 I,$$

$$E_2^T E_2 = \tilde{A}_{22}^T \tilde{A}_{22} + (\xi_0^{-2} \|\tilde{r}_1\| \|\tilde{y}_1\|)^2 I + (\theta^{-1} \xi_0^{-2} \|\tilde{r}_1\|)^2 I,$$

$$E^T f = Q \begin{pmatrix} E_1^T f_1 \\ 0 \end{pmatrix},$$

$$E_1 \equiv \begin{pmatrix} \tilde{A}_{11} + \xi_0^{-1} \xi_1 \tilde{r}_1 \tilde{y}_1^T \\ \xi_0^{-2} \|\tilde{r}_1\| \|\tilde{y}_1\| (I - \tilde{y}_1 \tilde{y}_1^\dagger) \\ \theta^{-1} \xi_0^{-2} \|\tilde{r}_1\| I \end{pmatrix},$$

$$f_1 \equiv \begin{pmatrix} \frac{\theta}{\sqrt{1+\theta^2 \|\tilde{y}_1\|^2}} \tilde{r} \\ 0 \\ \frac{\|\tilde{r}_1\|(\theta^2 - \gamma^2)}{(1+\gamma^2 \|\tilde{y}_1\|^2) \sqrt{1+\theta^2 \|\tilde{y}_1\|^2}} \tilde{y}_1 \end{pmatrix}.$$

It follows that

$$E(E^T E)^{-1} E^T f = \begin{pmatrix} P & & \\ & Q & \\ & & Q \end{pmatrix} \begin{pmatrix} (\tilde{A}_{11} + \xi_0^{-1} \xi_1 \tilde{r}_1 \tilde{y}_1^T) \\ 0 \\ \xi_0^{-2} \|\tilde{r}_1\| \|\tilde{y}_1\| (I - \tilde{y}_1 \tilde{y}_1^\dagger) \\ 0 \\ \theta^{-1} \xi_0^{-2} \|\tilde{r}_1\| I \\ 0 \end{pmatrix} (E_1^T E_1)^{-1} E_1^T f_1.$$

In addition, we know

$$E_1(E_1^T E_1)^{-1} E_1^T f_1 = \begin{pmatrix} (\tilde{A}_{11} + \xi_0^{-1} \xi_1 \tilde{r}_1 \tilde{y}_1^T) \\ \xi_0^{-2} \|\tilde{r}_1\| \|\tilde{y}_1\| (I - \tilde{y}_1 \tilde{y}_1^\dagger) \\ \theta^{-1} \xi_0^{-2} \|\tilde{r}_1\| I \end{pmatrix} (E_1^T E_1)^{-1} E_1^T f_1.$$

Then, we find

$$\varsigma(\gamma, \gamma, \theta) = \|E(E^T E)^{-1} E^T c\| = \|E_1(E_1^T E_1)^{-1} E_1^T f_1\| = \varsigma(\tilde{y}_1, \gamma, \theta).$$

□

3. Lanczos bidiagonalization

Given the matrix A and the starting vector b , the Lanczos bidiagonalization algorithm [14] generates the sequences of vectors u_i , v_i and scalars α_i , β_i satisfying the recurrences

$$\left. \begin{aligned} \beta_1 u_1 &= b, & \alpha_1 v_1 &= A^T u_1, \\ \beta_{k+1} u_{k+1} &= A v_k - \alpha_k u_k \\ \alpha_{k+1} v_{k+1} &= A^T u_{k+1} - \beta_{k+1} v_k \end{aligned} \right\}, \quad k = 1, 2, \dots$$

The scalars $\alpha_i \geq 0$ and $\beta_i \geq 0$ are chosen so that $\|u_i\|_2 = \|v_i\|_2 = 1$. With the definitions

$$U_k = \begin{pmatrix} u_1 & u_2 & \cdots & u_k \end{pmatrix}, \quad V_k = \begin{pmatrix} v_1 & v_2 & \cdots & v_k \end{pmatrix},$$

$$B_k \equiv \begin{pmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \beta_3 & \ddots & \\ & & \ddots & \alpha_k \\ & & & \beta_{k+1} \end{pmatrix}, \quad \bar{B}_k \equiv \begin{pmatrix} B_k & \alpha_{k+1} e_{k+1} \end{pmatrix}, \quad (3.1)$$

the recurrence relations (3.1) may be rewritten as

$$\begin{aligned} U_{k+1} \beta_1 e_1 &= b, \\ A V_k &= U_{k+1} B_k, \\ A^T U_{k+1} &= V_k B_k^T + \alpha_{k+1} v_{k+1} e_{k+1}^T = V_{k+1} \bar{B}_k^T. \end{aligned} \quad (3.2)$$

If exact arithmetic were used, we would also have $U_{k+1}^T U_{k+1} = I$, $V_k^T V_k = I$; in any event, the preceding equations hold to within machine precision. Thus, after k Lanczos iterations, we can project the STLS problem onto the subspaces spanned by U_k and V_k in the hope that, for sufficiently large k , we have captured all the large singular values of A needed to compute a useful regularized solution. The projected STLS problem is equivalent to

$$\min \|U_{k+1}^T ((b, A) - (\hat{b}_k, \hat{A}_k)) \begin{pmatrix} 1 & 0 \\ 0 & V_k \end{pmatrix}\|_F \quad \text{s.t.} \quad U_{k+1}^T \hat{A}_k V_k y = U_{k+1}^T \hat{b}_k,$$

or

$$\min \|(\beta_1 e_1, B_k) - (\hat{e}_k, \hat{B}_k)\|_F \quad \text{s.t.} \quad \hat{B}_k y = \hat{e}_k, \quad (3.3)$$

where $e_1 = (1, 0, \dots, 0)^T$, and $\hat{B}_k$ and $\hat{e}_k$ are generally full [15]. In each Lanczos step, we can now compute an approximate STLS solution x_k by applying the Algorithm TLS to the small-size problem in (3.3). This can be done in $O(k^2)$ operations by means of the partial singular value decomposition (PSVD) algorithm [16] or by methods posed in [17–19], and the approximate TLS solution x_k is given by $x_k = V_k y_k$, where y_k is the STLS solution to (3.3).

According to Paige [5], B_k and $\beta_1 e_1$ always satisfy (1.4), so the scaled TLS solution to any core problem $B_k y_k \approx \beta_1 e_1$ exists and is unique. The bidiagonalization removes all redundancies and irrelevant components of the problem associated with all singular values implicitly, prior to any singular value computation. Minimizing the dimensions of $(\beta_1 e_1, B_k)$ serves to maximize computational efficiency.

4. Estimating $\eta(x_k, \gamma, \theta)$

In this section, we present new bounds on and estimates of $\eta(x_k, \gamma, \theta)$ and its asymptotic estimate that can be estimated efficiently at every iteration of Lanczos bidiagonalization.

4.1. Practical bounds on $\eta(x_k, \gamma, \theta)$

We can approximate the value of $\eta(x_k, \gamma, \theta)$ by computing the backward error defined in Lemma 2.1 for the vector y associated with the projected TLS problem (3.3). In what follows, we show that for any integer $d \geq 0$, the choices $U = U_{k+d+1}$ and $V = V_{k+d}$ yield a lower bound on $\eta(x_k, \gamma, \theta)$, whereas $U = U_{k+d+1}$ and $V = V_{k+d+1}$ yield an upper bound. In the bases V_{k+d} and V_{k+d+1} , the vector x_k has coordinates

$$x_k = V_{k+d} \begin{pmatrix} y_k \\ 0_d \end{pmatrix} = V_{k+d} \hat{y}_k, \quad x_k = V_{k+d+1} \begin{pmatrix} y_k \\ 0_{d+1} \end{pmatrix} = V_{k+d} \bar{y}_k$$

with $r_k = b - Ax_k$. We define

$$\underline{\eta}_d(x_k, \gamma, \theta) \equiv \min_{\Delta B, \Delta c \in S_{B_{k+d}, \beta_1 e_1}^+(\gamma)} \|(\Delta B, \Delta c \theta)\|_F, \quad (4.1a)$$

$$\bar{\eta}_d(x_k, \gamma, \theta) \equiv \min_{\Delta B, \Delta c \in S_{\bar{B}_{k+d}, \beta_1 e_1}^+(\gamma)} \|(\Delta B, \Delta c \theta)\|_F \quad (4.1b)$$

with

$$\begin{aligned} S_{B_{k+d}, \beta_1 e_1}^+(\gamma) &\equiv \{(\Delta B, \Delta c) : (B_{k+d} + \Delta B)^T(\beta_1 e_1 + \Delta c - (B_{k+d} + \Delta B)\hat{y}_k) \\ &= -\frac{\|\beta_1 e_1 + \Delta c - (B_{k+d} + \Delta B)\hat{y}_k\|^2}{\gamma^{-2} + \|\hat{y}_k\|^2} \hat{y}_k\}, \\ S_{\bar{B}_{k+d}, \beta_1 e_1}^+(\gamma) &\equiv \{(\Delta B, \Delta c) : (\bar{B}_{k+d} + \Delta B)^T(\beta_1 e_1 + \Delta c - (\bar{B}_{k+d} + \Delta B)\bar{y}_k) \\ &= -\frac{\|\beta_1 e_1 + \Delta c - (\bar{B}_{k+d} + \Delta B)\bar{y}_k\|^2}{\gamma^{-2} + \|\bar{y}_k\|^2} \bar{y}_k\}, \end{aligned}$$

respectively. Letting

$$\hat{t}_k = \beta_1 e_1 - B_{k+d} \hat{y}_k = \beta_1 e_1 - \bar{B}_{k+d} \bar{y}_k, \quad (4.2)$$

we obtain

$$\underline{\eta}_d(x_k, \gamma, \theta) = \begin{cases} \omega_k^2, & \text{if } \lambda_{\min}(M_{k,d}) \geq 0; \\ \omega_k^2 + \lambda_{\min}(M_{k,d}) = \sigma_{\min}^2(N_{k,d}), & \text{if } \lambda_{\min}(M_{k,d}) < 0. \end{cases} \quad (4.3a)$$

$$\bar{\eta}_d(x_k, \gamma, \theta) = \begin{cases} \omega_k^2, & \text{if } \lambda_{\min}(\bar{M}_{k,d}) \geq 0; \\ \omega_k^2 + \lambda_{\min}(\bar{M}_{k,d}) = \sigma_{\min}^2(\bar{N}_{k,d}), & \text{if } \lambda_{\min}(\bar{M}_{k,d}) < 0, \end{cases} \quad (4.3b)$$

where

$$N_{k,d} \equiv \begin{pmatrix} B_{k+d}(I - \hat{y}_k \hat{y}_k^\dagger) & \omega_k(I - \hat{t}_k \hat{t}_k^\dagger) & v_k(B_{k+d} \hat{y}_k + \gamma^2 \|\hat{y}_k\|^2 \beta_1 e_1) \end{pmatrix}, \quad (4.4)$$

$$\begin{aligned} M_{k,d} &\equiv M(y, \gamma, \theta) \equiv B_{k+d}(I - \hat{y}_k \hat{y}_k^\dagger) B_{k+d}^T + \frac{\theta^2}{1 + \theta^2 \|\hat{y}_k\|^2} \hat{t}_k \hat{t}_k^T \\ &+ v_k^2(B_{k+d} \hat{y}_k + \gamma^2 \|\hat{y}_k\|^2 \beta_1 e_1) \times (B_{k+d} \hat{y}_k + \gamma^2 \|\hat{y}_k\|^2 \beta_1 e_1)^T = N_{k,d} N_{k,d}^T - \omega_k^2 I, \end{aligned}$$

$$\bar{N}_{k,d} \equiv \begin{pmatrix} \bar{B}_{k+d}(I - \bar{y}_k \bar{y}_k^\dagger) & \omega_k(I - \hat{t}_k \hat{t}_k^\dagger) & v_k(\bar{B}_{k+d} \bar{y}_k + \gamma^2 \|\bar{y}_k\|^2 \beta_1 e_1) \end{pmatrix}, \quad (4.5)$$

$$\begin{aligned}\bar{M}_{k,d} &\equiv M(y, \gamma, \theta) \equiv \bar{B}_{k+d}(I - \bar{y}_k \bar{y}_k^\dagger) \bar{B}_{k+d}^T + \frac{\theta^2}{1 + \theta^2 \|\bar{y}_k\|^2} \hat{t}_k \hat{t}_k^T \\ &\quad + v_k^2 (\bar{B}_{k+d} \bar{y}_k + \gamma^2 \|\bar{y}_k\|^2 \beta_1 e_1) \times (\bar{B}_{k+d} \bar{y}_k + \gamma^2 \|\bar{y}_k\|^2 \beta_1 e_1)^T = \bar{N}_{k,d} \bar{N}_{k,d}^T - \omega_k^2 I,\end{aligned}$$

where $\omega_k = \theta \|r_k\| / \sqrt{1 + \theta^2 \|y_k\|^2}$, $v_k = \theta / \sqrt{\theta^2 \|y_k\|^2 + \gamma^4 \|y_k\|^4}$.

Theorem 4.1. Let $x_k = V_k y_k$ be the approximate solution to the scaled total least squares problem (1.3), computed by bidiagonalization at iteration step k , and let $\eta(x_k, \gamma, \theta)$ be the extended backward error associated with x_k for given $\gamma > 0$, $\theta > 0$. Let $\underline{\eta}_d(x_k, \gamma, \theta)$ and $\bar{\eta}_d(x_k, \gamma, \theta)$ be defined by (4.3a) and (4.3b) for a given $d \geq 0$. Then,

$$\underline{\eta}_d(x_k, \gamma, \theta) \leq \eta(x_k, \gamma, \theta) \leq \bar{\eta}_d(x_k, \gamma, \theta). \quad (4.6)$$

In addition, the left and right equations are established for some iteration step $s \leq n - k$.

Proof. If $\lambda_{\min}(M) \geq 0$, then $\underline{\eta}_d(x_k, \gamma, \theta) \leq \eta(x_k, \gamma, \theta)$ holds.

On the other hand assume, that $\lambda_{\min}(M) < 0$; that is, $\eta(x_k, \gamma, \theta) = \sigma_{\min}(N)$. There are matrices $\tilde{U}$ and $\tilde{V}$ such that $U = (U_{k+d+1}, \tilde{U})$ and $V = (V_{k+d}, \tilde{V})$ are orthogonal matrices. Then, from (3.2) and (4.2),

$$\begin{aligned}U^T N \begin{pmatrix} v & u \\ & I \end{pmatrix} &= \begin{pmatrix} U_{k+d+1}^T \\ \tilde{U}^T \end{pmatrix} \begin{pmatrix} A(I - yy^\dagger) & \omega_k(I - rr^\dagger) & v_k(Ay + \gamma^2 \|y\|^2 b) \\ C & D & 0 \end{pmatrix} \\ &\quad \times \begin{pmatrix} V_{k+d} & \tilde{V} & 0 & 0 & 0 \\ 0 & 0 & U_{k+d+1} & \tilde{U} & 0 \\ 0 & 0 & 0 & 0 & I \end{pmatrix}, \\ &= \begin{pmatrix} B_{k+d}(I - \hat{y}_k \hat{y}_k^\dagger) & C & \omega_k(I - \hat{t}_k \hat{t}_k^\dagger) & 0 & v_k(B_{k+d} \hat{y}_k + \gamma^2 \|\hat{y}_k\|^2 \beta_1 e_1) \\ 0 & D & 0 & \omega_k I & 0 \end{pmatrix},\end{aligned} \quad (4.7)$$

where we use the following results:

$$\begin{aligned}\tilde{U}^T A(I - yy^\dagger) V_{k+d} &= 0, \\ \tilde{U}^T \omega_k(I - rr^\dagger) U_{k+d+1} &= 0, \\ v_k \tilde{U}^T (Ay + \gamma^2 \|y\|^2 b) &= 0, \\ C &= U_{k+d+1}^T A(I - yy^\dagger) \tilde{V} = U_{k+d+1}^T A \tilde{V}, \\ D &= \tilde{U}^T A(I - yy^\dagger) \tilde{V} = \tilde{U}^T A \tilde{V}.\end{aligned}$$

Owing to the interlacing property of singular values (see, e.g., [9]) the smallest singular value of N is not less than the smallest singular value of the matrix

$$\begin{pmatrix} B_{k+d}(I - \hat{y}_k \hat{y}_k^\dagger) & \omega_k(I - \hat{t}_k \hat{t}_k^\dagger) & v_k(B_{k+d} \hat{y}_k + \gamma^2 \|\hat{y}_k\|^2 \beta_1 e_1) & 0 \\ 0 & 0 & 0 & \omega_k I \end{pmatrix} = \begin{pmatrix} N_{k,d} & 0 \\ 0 & \omega_k I \end{pmatrix}$$

obtained from (4.7) after deleting the columns corresponding to the matrices C and D and making column permutation. Hence,

$$\eta(x_k, \gamma, \theta) = \sigma_{\min}(N) \geq \sigma_{\min}(N_{k+d}) = \underline{\eta}_d(x_k, \gamma, \theta)$$

due to the assumption $\sigma_{\min}(N) \leq \omega_k$. The lower bound of $\eta(x_k, \gamma, \theta)$ also comes from the interleaving property. Because y_k is the exact solution of the total least squares problem (3.3), $\underline{\eta}_0(x_k, \gamma, \theta) = 0$ is obtained, and the existence of s makes $\underline{\eta}_s(x_k, \gamma, \theta) = \eta(x_k, \gamma, \theta)$ determined by the finite termination property of bidiagonalization.

For the upper bound in (4.6), we have

$$\begin{aligned} \sigma_{\min}(N) &= \min_{q: \|q\|_2=1} \left\| \begin{pmatrix} (I - yy^\dagger)A^T \\ \omega_k(I - rr^\dagger) \\ v_k(Ay + \gamma^2\|y\|^2b)^T \end{pmatrix} q \right\|_2 \\ &\leq \min_{\tilde{q}: \|\tilde{q}\|_2=1} \left\| \begin{pmatrix} (I - yy^\dagger)A^T \\ \omega_k(I - rr^\dagger) \\ v_k(Ay + \gamma^2\|y\|^2b)^T \end{pmatrix} U_{k+d+1} \tilde{q} \right\|_2 \\ &= \min_{\tilde{q}: \|\tilde{q}\|_2=1} \left\| \begin{pmatrix} V_{k+d+1}(I - \bar{y}_k \bar{y}_k^\dagger) \bar{B}_{k+d}^T \\ U_{k+d+1} \omega_k(I - \hat{t}_k \hat{t}_k^\dagger) \\ v_k(\bar{B}_{k+d} \bar{y}_k + \gamma^2\|\bar{y}_k\|^2 \beta_1 e_1)^T \end{pmatrix} \tilde{q} \right\|_2 \\ &= \min_{\tilde{q}: \|\tilde{q}\|_2=1} \left\| \begin{pmatrix} (I - \bar{y}_k \bar{y}_k^\dagger) \bar{B}_{k+d}^T \\ \omega_k(I - \hat{t}_k \hat{t}_k^\dagger) \\ v_k(\bar{B}_{k+d} \bar{y}_k + \gamma^2\|\bar{y}_k\|^2 \beta_1 e_1)^T \end{pmatrix} \tilde{q} \right\|_2 = \sigma_{\min}(\bar{N}_{k,d}). \end{aligned}$$

Hence, it follows that $\eta(x_k, \gamma, \theta) \leq \bar{\eta}_d(x_k, \gamma, \theta)$. The upper bound (4.6) comes from the fact that as d grows, we are expanding the subspace $R(U_{k+d+1})$, which we are minimizing to obtain the upper bound. Because $\bar{N}_{k,s}$ and $N_{k,s}$ differ in only one zero column, we get the equivalent $\eta(x_k, \gamma, \theta) = \bar{\eta}_s(x_k, \gamma, \theta)$. $\square$

To evaluate the estimates $\underline{\eta}_d(x_k, \gamma, \theta)$ and $\bar{\eta}_d(x_k, \gamma, \theta)$ in (4.3a) and (4.3b), we should compute the smallest singular value of the matrices $N_{k,d}$ and $\bar{N}_{k,d}$ in (4.4) and (4.5). If we let $\bar{Q}_k = \text{diag}(Q_k, I_d)$, Q_k is the matrix of k reflections from (4.10). Then, from (4.2), we know the matrix $N_{k,d}$ is orthogonally similar to the matrix

$$L_{k,d} \equiv \bar{Q}_k N_{k,d} \begin{pmatrix} I_{k,d} & & \\ & \bar{Q}_k^T & \\ & & 1 \end{pmatrix} = \begin{pmatrix} \bar{Q}_k B_{k+d} & \omega(I - e_{k+1} e_{k+1}^T) & \bar{Q}_k v(B_{k+d} \hat{y}_k + \gamma^2\|\hat{y}_k\|^2 \beta_1 e_1) \end{pmatrix}.$$

The matrix $\bar{Q}_k B_{k+d}$ is the transformed matrix B_{k+d} partially. In the same way, the computation of $\bar{\eta}_d(x_k, \gamma, \theta)$ can be treated. $\bar{N}_{k,d}$ is orthogonally similar to the matrix

$$\begin{aligned} \bar{L}_{k,d} &\equiv \bar{Q}_k \bar{N}_{k,d} \begin{pmatrix} I_{k+d+1} & & \\ & \bar{Q}_k^T & \\ & & 1 \end{pmatrix} = \begin{pmatrix} \bar{Q}_k \bar{B}_{k+d} & \omega(I - e_{k+1} e_{k+1}^T) & \bar{Q}_k v(B_{k+d} \hat{y}_k + \gamma^2\|\hat{y}_k\|^2 \beta_1 e_1) \end{pmatrix} \\ &= \begin{pmatrix} \bar{Q}_k B_{k+d} & \alpha_{k+d+1} e_{k+d+1} & \omega(I - e_{k+1} e_{k+1}^T) & \bar{Q}_k v(B_{k+d} \hat{y}_k + \gamma^2\|\hat{y}_k\|^2 \beta_1 e_1) \end{pmatrix}, \end{aligned}$$

which is different from $L_{k,d}$ by just one column, $\alpha_{k+d+1} e_{k+d+1}$.

Because of the special structure of $L_{k,d}$ and $\bar{L}_{k,d}$, the smallest singular values of $L_{k,d}$ and $\bar{L}_{k,d}$ can be computed very efficiently. The implementation details can be found in the Appendix.

4.2. An asymptotic estimate of $\eta(x_k, \gamma, \theta)$

We know that Chang [7] give an asymptotic estimate $\varsigma(x_k, \gamma, \theta)$ for $\eta(x_k, \gamma, \theta)$ just as (2.8). $\varsigma(x_k, \gamma, \theta)$ and $\eta(x_k, \gamma, \theta)$ are asymptotically equivalent:

$$\lim_{x_k \rightarrow x} \frac{\varsigma(x_k, \gamma, \theta)}{\eta(x_k, \gamma, \theta)} = 1,$$

where x is the true solution of the STLS problem (1.3). Just as posed in [13], we consider the asymptotic estimates associated with the backward errors $\underline{\eta}_d(x_k, \gamma, \theta)$ and $\bar{\eta}_d(x_k, \gamma, \theta)$ in (4.3a) and (4.3b), namely

$$\underline{\varsigma}_d(x_k, \gamma, \theta) = \|F_1 F_1^\dagger g_1\|, \quad (4.8)$$

$$\bar{\varsigma}_d(x_k, \gamma, \theta) = \|F_2 F_2^\dagger g_2\|, \quad (4.9)$$

where

$$F_1 \equiv \begin{pmatrix} B_{k+d} + \xi_0^{-1} \xi_1 \hat{t}_k \hat{y}_k^T \\ \xi_0^{-2} \|\hat{t}_k\| \|\hat{y}_k\| (I - \hat{y}_k \hat{y}_k^\dagger) \\ \theta^{-1} \xi_0^{-2} \|\hat{t}_k\| I \end{pmatrix}, g_1 \equiv \begin{pmatrix} \frac{\theta}{\sqrt{1+\theta^2 \|\hat{y}_k\|^2}} \hat{t}_k \\ 0 \\ \frac{\|\hat{t}_k\|(\theta^2 - \gamma^2)}{(1+\gamma^2 \|\hat{y}_k\|^2) \sqrt{1+\theta^2 \|\hat{y}_k\|^2}} \hat{y}_k \end{pmatrix},$$

$$F_2 \equiv \begin{pmatrix} \bar{B}_{k+d} + \xi_0^{-1} \xi_1 \hat{t}_k \bar{y}_k^T \\ \xi_0^{-2} \|\hat{t}_k\| \|\bar{y}_k\| (I - \bar{y}_k \bar{y}_k^\dagger) \\ \theta^{-1} \xi_0^{-2} \|\hat{t}_k\| I \end{pmatrix}, g_2 \equiv \begin{pmatrix} \frac{\theta}{\sqrt{1+\theta^2 \|\bar{y}_k\|^2}} \hat{t}_k \\ 0 \\ \frac{\|\hat{t}_k\|(\theta^2 - \gamma^2)}{(1+\gamma^2 \|\bar{y}_k\|^2) \sqrt{1+\theta^2 \|\bar{y}_k\|^2}} \bar{y}_k \end{pmatrix}.$$

Transform B_k to upper bidiagonal form using a product of reflections Q_k , leading to

$$Q_k \begin{pmatrix} B_k & \beta_1 e_1 \end{pmatrix} = \begin{pmatrix} R_k & f_k \\ 0 & \bar{\phi}_{k+1} \end{pmatrix}, \quad (4.10)$$

$$R_k \equiv \begin{pmatrix} \rho_1 & \theta_2 & & \\ & \rho_2 & \ddots & \\ & & \ddots & \theta_k \\ & & & \rho_k \end{pmatrix}, \quad f_k \equiv \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_k \end{pmatrix}.$$

Using the QR decomposition of B_{k+d} from (4.10) and its QR decomposition combined with rank-1 correction, we obtain

$$\begin{pmatrix} G_1 Q_{k+d} \\ G_2 \\ I \end{pmatrix} \begin{pmatrix} B_{k+d} + \xi_0^{-1} \xi_1 \hat{t}_k \hat{y}_k^T \\ \xi_0^{-2} \|\hat{t}_k\| \|\hat{y}_k\| (I - \hat{y}_k \hat{y}_k^T) \\ \theta^{-1} \xi_0^{-2} \|\hat{t}_k\| I \end{pmatrix} = \begin{pmatrix} * & * & \cdots & * \\ & * & \cdots & * \\ & & \ddots & \vdots \\ & & & * \\ 0 & 0 & \cdots & 0 \\ * & * & \cdots & * \\ & * & \cdots & * \\ & & \ddots & \vdots \\ & & & * \\ * & & & \\ & * & & \\ & & \ddots & \\ & & & * \end{pmatrix} \equiv \begin{pmatrix} C_{k+d}^1 \\ C_{k+d}^2 \\ C_{k+d}^3 \end{pmatrix} \equiv C_{k+d}.$$

We can eliminate the $\theta^{-1} \xi_0^{-2} \|\hat{t}_k\| I$ block in C_{k+d} using a product, denoted by G_{k+d} , of $2(k+d) - 1$ Givens rotations. Here is a schematic illustration of the first two such Givens rotations in the case of $k+d=4$:

$$\begin{pmatrix} * & * & * & * \\ & * & * & * \\ & & * & * \\ & & & * \\ 0 & 0 & 0 & 0 \\ \bullet & \bullet & \bullet & \bullet \\ & * & * & * \\ & & * & * \\ & & & * \\ 1 \odot & 2 \odot & & \\ & \bullet & & * \\ & & & * \end{pmatrix}$$

For submatrix C_{k+d}^2 in C_{k+d} , we can eliminate it column by column using a product of $(k+d)(k+d+1)/2$ Givens rotations as follows, which we denote $\bar{G}_{k+d}$.

$$\begin{pmatrix}
\bullet & \bullet & \bullet & \bullet \\
& * & * & * \\
& & * & * \\
& & & * \\
0 & 0 & 0 & 0 \\
\odot & \bullet & \bullet & \bullet \\
& * & * & * \\
& & * & * \\
& & & * \\
0 & & & \\
& 0 & & \\
& & 0 & \\
& & & 0
\end{pmatrix}
\quad
\begin{pmatrix}
* & * & * & * \\
& \bullet & \bullet & \bullet \\
& & * & * \\
& & & * \\
0 & 0 & 0 & 0 \\
0 & \odot & \bullet & \bullet \\
& \odot & \bullet & \bullet \\
& & * & * \\
& & & * \\
0 & & & \\
& 0 & & \\
& & 0 & \\
& & & 0
\end{pmatrix}$$

$$\begin{pmatrix}
* & * & * & * \\
& * & * & * \\
& & \bullet & \bullet \\
0 & 0 & 0 & 0 \\
0 & 0 & \odot & \bullet \\
& 0 & \odot & \bullet \\
& & \odot & \bullet \\
& & & * \\
0 & & & \\
& 0 & & \\
& & 0 & \\
& & & 0
\end{pmatrix}
\quad
\begin{pmatrix}
* & * & * & * \\
& * & * & * \\
& & * & * \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \odot \\
& 0 & 0 & \odot \\
& & 0 & \odot \\
& & & \odot \\
0 & & & \\
& 0 & & \\
& & 0 & \\
& & & 0
\end{pmatrix}$$

After the entire C_{k+d}^2 and C_{k+d}^3 block of C_{k+d} has been zeroed in this way, we have

$$G_{k+d}C_{k+d} = G_{k+d} \begin{pmatrix} G_1 Q_{k+d} & \\ & G_2 \\ & & I \end{pmatrix} \begin{pmatrix} B_{k+d} + \xi_0^{-1} \xi_1 \hat{t}_k \hat{y}_k^T \\ \xi_0^{-2} \|\hat{t}_k\| \|\hat{y}_k\| (I - \hat{y}_k \hat{y}_k^T) \\ \theta^{-1} \xi_0^{-2} \|\hat{t}_k\| I \end{pmatrix} = \begin{pmatrix} \tilde{R}_{k+d} \\ 0 \\ 0 \end{pmatrix},$$

where $\tilde{R}_{k+d}$ is upper triangular, from which it follows that

$$\begin{pmatrix} B_{k+d} + \xi_0^{-1} \xi_1 \hat{t}_k \hat{y}_k^T \\ \xi_0^{-2} \|\hat{t}_k\| \|\hat{y}_k\| (I - \hat{y}_k \hat{y}_k^T) \\ \theta^{-1} \xi_0^{-2} \|\hat{t}_k\| I \end{pmatrix}^\dagger = \begin{pmatrix} \tilde{R}_{k+d}^{-1} & 0 & 0 \end{pmatrix} G_{k+d} \begin{pmatrix} G_1 Q_{k+d} & \\ & G_2 \\ & & I \end{pmatrix},$$

and, we obtain

$$F_1 F_1^\dagger g_1 = \begin{pmatrix} (G_1 Q_{k+d})^T \\ G_2^T \\ I \end{pmatrix} G_{k+d}^T \begin{pmatrix} I_{k+d} & \\ & 0 \\ & & 0 \end{pmatrix} G_{k+d} \begin{pmatrix} G_1 Q_{k+d} & \\ & G_2 \\ & & I \end{pmatrix} \begin{pmatrix} \frac{\theta}{\sqrt{1+\theta^2 \|\hat{y}_k\|^2}} \hat{t}_k \\ 0 \\ \frac{\|\hat{t}_k\| (\theta^2 - \gamma^2)}{(1+\gamma^2 \|\hat{y}_k\|^2) \sqrt{1+\theta^2 \|\hat{y}_k\|^2}} \hat{y}_k \end{pmatrix}.$$

Taking the 2-norm, we obtain

$$\underline{s}_d(x_k, \gamma, \theta) = \|F_1 F_1^\dagger g_1\| = \left\| \begin{pmatrix} I_{k+d} & & \\ & 0 & \\ & & 0 \end{pmatrix} G_{k+d} \begin{pmatrix} G_1 & & \\ & G_2 & \\ & & I \end{pmatrix} \begin{pmatrix} s_k \\ 0 \\ \zeta \hat{y}_k \end{pmatrix} \right\|,$$

where $s_k = Q_{k+d} \hat{t}_k = \begin{pmatrix} 0 & \cdots & 0 & \phi_{k+1} & \cdots & \phi_{k+d} & \bar{\phi}_{k+d+1} \end{pmatrix}^T$; see [13, Equation (4.6)] and $\zeta = \|\hat{t}_k\|(\theta^2 - \gamma^2)/[(1 + \gamma^2\|\hat{y}_k\|^2)\sqrt{1 + \theta^2\|\hat{y}_k\|^2}]$. By applying $26(k+d)^2$ Givens rotations to the vector s_k and $[(k+d)(k+d+5) - 1]/2$ Givens rotations to the vector $(s_k, 0, \zeta \hat{y}_k)^T$, the estimate $\eta_{k+d}(x_k, \gamma, \theta)$ can be computed, along with computing the 2-norm of the vector formed by its first $k+d$ elements.

The estimate $\bar{s}_d(x_k, \gamma, \theta)$ can be computed in a similar way. Instead of (4.11), we have

$$\begin{pmatrix} G_3 Q_{k+d} & \\ & G_4 \\ & & I \end{pmatrix} \begin{pmatrix} \bar{B}_{k+d} + \xi_0^{-1} \xi_1 \hat{t}_k \bar{y}_k^T \\ \xi_0^{-2} \|\hat{t}_k\| \|\bar{y}_k\| (I - \bar{y}_k \bar{y}_k^\dagger) \\ \theta^{-1} \xi_0^{-2} \|\hat{t}_k\| I \end{pmatrix} = \begin{pmatrix} * & * & \cdots & * & * \\ & * & \cdots & * & * \\ & & \ddots & \vdots & \vdots \\ & & & * & * \\ 0 & 0 & \cdots & 0 & * \\ * & * & \cdots & * & * \\ & * & \cdots & * & * \\ & & \ddots & \vdots & \vdots \\ & & & * & * \\ * & & & & * \\ & * & & & * \\ & & \ddots & & \\ & & & * & * \end{pmatrix},$$

which differs from C_{k+d} by only one extra column. Thus, by using $k+d+3$ extra Givens rotations, $\bar{s}_d(x_k, \gamma, \theta)$ can be computed in much the same way as $\underline{s}_d(x_k, \gamma, \theta)$.

Note: STLS unifies OLS, TLS, and DLS. The minimal backward error estimates for STLS derived in this paper also unify the backward perturbation analyses of OLS, TLS, and DLS. Thus, this work unifies and generalizes the results in [13], in which practical and inexpensive computable estimates of the backward error for the least squares problem are presented.

5. Estimating stopping criterion ζ_k efficiently

Let the SVD of (b, A) be $W \text{diag}(\sigma_i) Z^T$, and define $P_{U_k}^\perp \equiv (I - U_k U_k^T)$ as the orthogonal projector onto the subspace $\text{span}\{U_k\}^\perp$, the orthogonal complement of $\text{span}\{U_k\}$. For convenience, define the matrices $\bar{U}^{(k)} \in R^{m \times (k+1)}$, $\bar{V}^{(k)} \in R^{n \times (k+1)}$, and $\bar{V}_k \in R^{(n+1) \times (k+2)}$ such that $\bar{U}^{(k)} = U_{k+1} W$, $\bar{V}^{(k)} = \bar{V}_k Z$ and $\bar{V}_k = \begin{pmatrix} 1 & 0 \\ 0 & V_k \end{pmatrix}$.

Now, if one is only interested in finding a low-rank approximation of (b, A) , it is desirable that a more direct approach be used without computing the singular value decomposition (SVD) of

$(\beta_1 e_1, B_k)$. From the convergence theory of the Lanczos process, we know that U_k and V_k contain good approximations of the dominant singular vectors of (b, A) . It is therefore natural to use $J_k \equiv U_k(\beta_1 e_1, \bar{B}_{k-1})\bar{V}_k^T$ as an approximation of (b, A) . In the next section, we consider an a priori estimate of $\zeta_k \equiv \|(b, A) - J_k\|_F$ and begin with a lemma.

Lemma 5.1. [20] Let $C \in R^{n \times n}$ be symmetric and f an arbitrary vector. Define $\mathcal{K}_k \equiv \text{span}\{f, Cf, \dots, C^{k-1}f\}$, and let $C = X \text{diag}(\lambda_i)X^T$ be the eigendecomposition of C with $\lambda_1 \geq \dots \geq \lambda_n$ its eigenvalues. Write $X = [x_1, \dots, x_n]$, and define $\mathcal{X}_i = \text{span}\{x_1, \dots, x_i\}$. Then,

$$\tan\theta(x_i, \mathcal{K}_k) \leq \frac{\sin\theta(f, \mathcal{X}_i)K_i}{\cos\theta(f, x_i)T_{k-i}(1 + 2\gamma_i)},$$

where T_k is the Chebyshev polynomial, and the following hold:

$$\begin{cases} K_1 = 1, \\ K_i = \prod_{j=1}^{i-1} \frac{\sigma_j^2 - \sigma_n^2}{\sigma_j^2 - \sigma_i^2}, \quad \text{if } i \neq 1, \end{cases} \quad \text{and} \quad \gamma_i = \frac{\sigma_i^2 - \sigma_{i+1}^2}{\sigma_{i+1}^2 - \sigma_n^2}.$$

Theorem 5.1. Let $\mathcal{W}_i \equiv \text{span}\{w_1, w_2, \dots, w_i\}$. Assume that the Lanczos bidiagonalization process starts with b . Then, for any j with $1 < j < n$ and $k > j$,

$$\zeta_k^2 \leq \sum_{i=j+1}^n \sigma_i^2 + \sum_{i=1}^j \sigma_i^2 \left(\frac{\sin\theta(b, \mathcal{W}_i)K_i}{\cos\theta(b, w_i)T_{k-i}(1 + 2\gamma_i)} \right)^2. \quad (5.1)$$

Proof. From (3.2), we have $U_k^T A = \bar{B}_{k-1} V_k^T$. Hence,

$$\zeta_k^2 = \|(b, A) - U_k(\beta_1 e_1, \bar{B}_{k-1})\bar{V}_k^T\|_F^2 = \|(I - U_k U_k^T)(b, A)\|_F^2.$$

Using the SVD of $(b, A) = W \text{diag}(\sigma_i)Z^T$, one can verify that

$$\begin{aligned} \zeta_k^2 &= \|\sigma_1 P_{U_k}^\perp w_1, \sigma_2 P_{U_k}^\perp w_2, \dots, \sigma_{n+1} P_{U_k}^\perp w_{n+1}\|_F^2 \\ &\leq \sum_{i=j+1}^{n+1} \sigma_i^2 + \sum_{i=1}^j \sigma_i^2 \|P_{U_k}^\perp w_i\|^2. \end{aligned}$$

It is readily verified that $\text{span}\{U_k\} = \text{span}\{b, AA^T b, \dots, (AA^T)^{k-1}b\} \equiv \mathcal{K}_k$. Therefore,

$$\|P_{U_k}^\perp w_i\| = |\sin \angle(w_i, \mathcal{K}_k)| \leq |\tan \angle(w_i, \mathcal{K}_k)|.$$

Applying Lemma 5.1 completes the proof. $\square$

Notice that the square root of the first term on the right-hand side of (5.1), $(\sum_{i=j+1}^{n+1} \sigma_i^2)^{1/2}$, is the error of the best rank- j approximation of (b, A) in the Frobenius norm. For $k > j$, we usually have $\text{rank}(J_k) > j$. The a priori estimate in the above theorem shows that ζ_k^2 approaches $\|(b, A) - (b, A)_j\|_F^2$ as k becomes large, where $(b, A)_j$ denotes the best rank- j approximation of (b, A) . In many examples, even when k is only slightly larger than j , ζ_k^2 is already very close to $\|(b, A) - (b, A)_j\|_F^2$. The accuracy of using $J_k = U_k(\beta_1 e_1, \bar{B}_{k-1})\bar{V}_k^T$ as a low-rank approximation of (b, A) is measured by ζ_k^2 , which can

be used as a stopping criterion in the iterative Lanczos bidiagonalization process; that is, the iteration terminates when

$$\zeta_k^2 \leq \text{tol}, \quad (5.2)$$

where tol is a user-supplied tolerance. One can find an inexpensive way to compute ζ_k^2 for $k = 1, 2, \dots$ as follows:

Theorem 5.2. Let $\zeta_k = \|(b, A) - U_k(\beta_1 e_1, \bar{B}_{k-1})\bar{V}_k^T\|_F$. Then,

$$\zeta_k^2 = \zeta_{k+d}^2 + \sum_{j=k}^{k+d-1} (\alpha_j^2 + \beta_{j+1}^2), \quad (5.3)$$

where α_j and β_{j+1} are from (3.1).

Proof. We know that

$$\zeta_{k+d}^2 = \|(b, A) - U_{k+d}(\beta_1 e_1, \bar{B}_{k+d-1})\bar{V}_{k+d}^T\|_F^2 = \|(I - U_{k+d}U_{k+d}^T)(b, A)\|_F^2;$$

we therefore write $I - U_k U_k^T = I - U_{k+d} U_{k+d}^T + \sum_{j=k+1}^{k+d} u_j u_j^T$. Then, we obtain

$$\zeta_k^2 = \|(I - U_{k+d} U_{k+d}^T)(b, A)\|_F^2 + \sum_{j=k+1}^{k+d} \|u_j u_j^T (b, A)\|_F^2 = \zeta_{k+d}^2 + \sum_{j=k+1}^{k+d} \|(b, A)^T u_j\|^2.$$

The proof is completed by noticing that $\|(b, A)^T u_j\|^2 = \alpha_{j-1}^2 + \beta_j^2$ for $j = k+1, \dots, k+d$. □

By Theorem 5.2, we can estimate ζ_k as follows. Ignoring the ζ_{k+d} term in (5.3) gives

$$\zeta_k^2 \geq \sum_{j=k}^{k+d-1} (\alpha_j^2 + \beta_{j+1}^2) \equiv \phi_d^2(\alpha, \beta). \quad (5.4)$$

The hope is that

$$\zeta_k^2 \gg \zeta_{k+d}^2 \quad (5.5)$$

in (5.3), and thus, that the bound in (5.4) is tight. This is certainly reasonable for sufficiently large d , as in the Lanczos bidiagonalization TLS algorithm, ζ_k is strictly monotonically decreasing theoretically with k and converges to 0. Notice that the estimate $\phi_d(\alpha, \beta)$ can be computed at essentially no extra cost in the Lanczos bidiagonalization TLS algorithm, but d additional iterations have to be performed to estimate ζ_k . A larger value of d results in a longer such delay; however, it also produces a larger difference in (5.5) and thus, a tighter bound in (5.4).

The estimate $\phi_d(\alpha, \beta)$ of ζ_k allows us to use (5.2) as a practical stopping criterion in the Lanczos bidiagonalization TLS method.

6. Numerical examples

We present several numerical examples to illustrate our results. Recall from Section 2 that the extended minimal backward error $\eta(y, \gamma, \theta)$ defined in (2.2) is a lower bound for the true minimal backward error. Furthermore, it follows from Theorem 2.3 that if y is sufficiently close to the true solution x , then $\eta(y, \gamma, \theta)$ is in fact the true minimal backward error. We set $\gamma = \theta = 1$ in all subsequent examples.

All numerical experiments are performed using MATLAB R2015a on a 3.20 GHz Intel Core(TM) 2 Duo CPU P8700 processor running Windows 7. The MATLAB built-in functions `randn` and `rand` are used below; they generate matrices whose entries follow the standard normal distribution and the standard uniform distribution, respectively. Our test problems are constructed as follows:

Example 6.1. Let $A = \tilde{A}/\|\tilde{A}\|_F$, $\tilde{A} = \text{randn}(120, 80)$. $b = Ax$, where $x = (1, \dots, 1)^T$. and let $A = A + (1/(\sqrt{120 \times 80}))\delta_A \cdot \text{randn}(120, 80)$ and $b = b + (1/(\sqrt{120}))\delta_b \cdot \text{rand}(120, 1)$ with $\delta_A = \delta_b = 10^{-15}, 10^{-10}, 10^{-6}, 10^{-2}$.

The lower bounds $\underline{\eta}_d(x_k, \gamma, \theta)$ in Figure 1 and the upper bounds $\bar{\eta}_d(x_k, \gamma, \theta)$ in Figure 2 for the backward error $\eta(x_k, \gamma, \theta)$ are compared with the true values $\eta(x_k, \gamma, \theta)$ for different values of the parameter d . As predicted by Theorem 4.1, the quality of the estimates improves as d increases. In this example, we observe that both the upper and lower bounds of the backward error become increasingly accurate as d grows. Values of d as small as ($d = 7$) yield excellent results for well-conditioned problems, as illustrated by Test Problem 1 in Figures 1 and 2. It is also shown that the upper and lower bounds of the backward error provide better estimates for small noise problems.

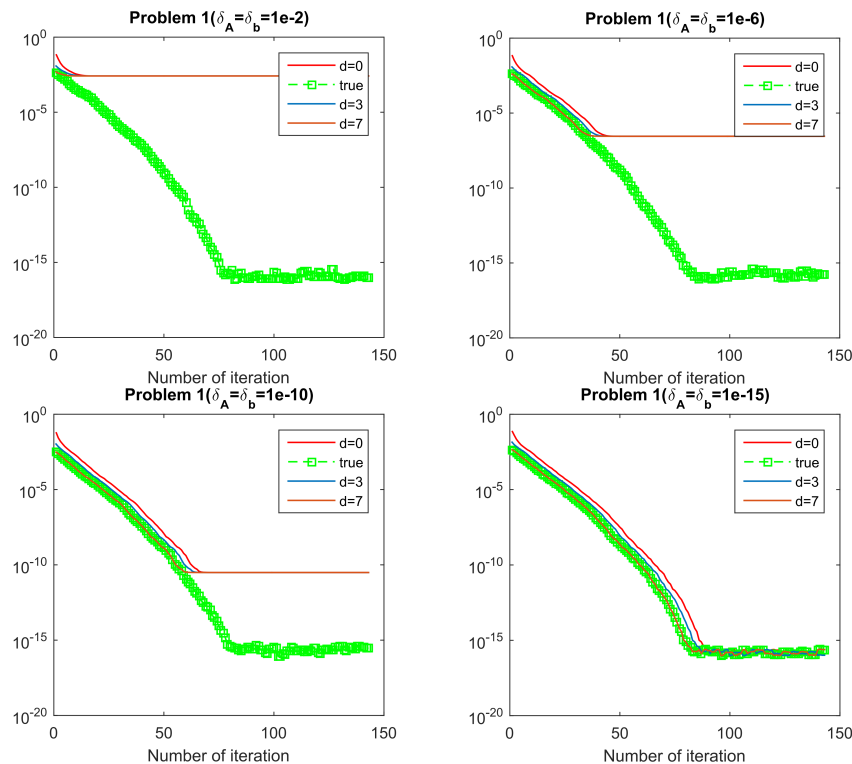


Figure 1. Estimating $\eta(x_k, \gamma, \theta)$ by using $\underline{\eta}_d(x_k, \gamma, \theta)$ with different d for a well-conditioned problem.

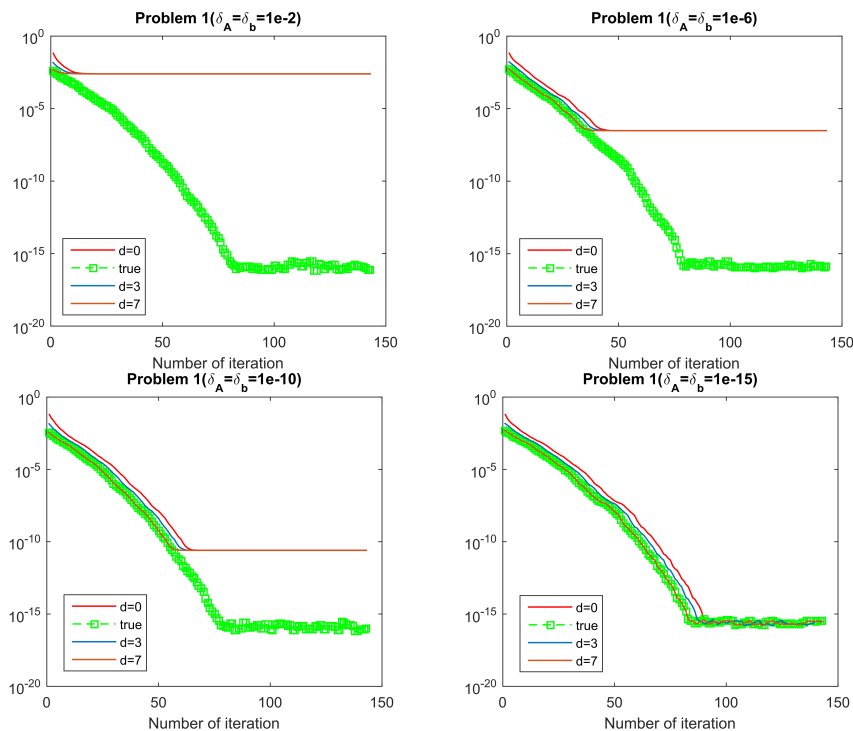


Figure 2. Estimating $\eta(x_k, \gamma, \theta)$ by using $\bar{\eta}_d(x_k, \gamma, \theta)$ with different d for a well-conditioned problem.

Example 6.2. Let $A = \tilde{A}/\|\tilde{A}\|_F$, $\tilde{A} = U\Sigma V^T$, where U and V are the Q -factors of the QR factorization of random matrices $\text{rand}(120, 120)$ and $\text{rand}(80, 80)$, respectively, and $\Sigma = \text{diag}(\sigma_i)$ with logarithmically equally spaced singular values σ_i between 1 and 10^{-4} . Here, $\kappa_2(A) = 10^{-4}$. $b = Ax$, where $x = (1, \dots, 1)^T$, and let $A = A + (1/(\sqrt{120 \times 80}))\delta_A \cdot \text{rand}(120, 80)$, and $b = b + (1/(\sqrt{120}))\delta_b \cdot \text{rand}(120, 1)$, with $\delta_A = \delta_b = 10^{-15}, 10^{-10}, 10^{-6}, 10^{-2}$.

The lower bounds $\underline{\eta}_d(x_k, \gamma, \theta)$ in Figure 3 and the upper bounds $\bar{\eta}_d(x_k, \gamma, \theta)$ in Figure 4 for the backward error $\eta(x_k, \gamma, \theta)$ are compared with the true values $\eta(x_k, \gamma, \theta)$ for various values of the parameter d . Similarly, we observe that as d increases, both the upper and lower bounds of the backward error provide increasingly accurate estimates. However, larger values of d are required compared to those for the well-conditioned problem. It is also shown that the upper and lower bounds of the backward error yield better estimates for problems with small noise. Values of d as small as $d = 9$ yield excellent results for ill-conditioned problems, as illustrated by Test Problem 2 in Figures 3 and 4. During the estimation process, the lower bound exhibits some volatility, whereas the upper bound is more stable.

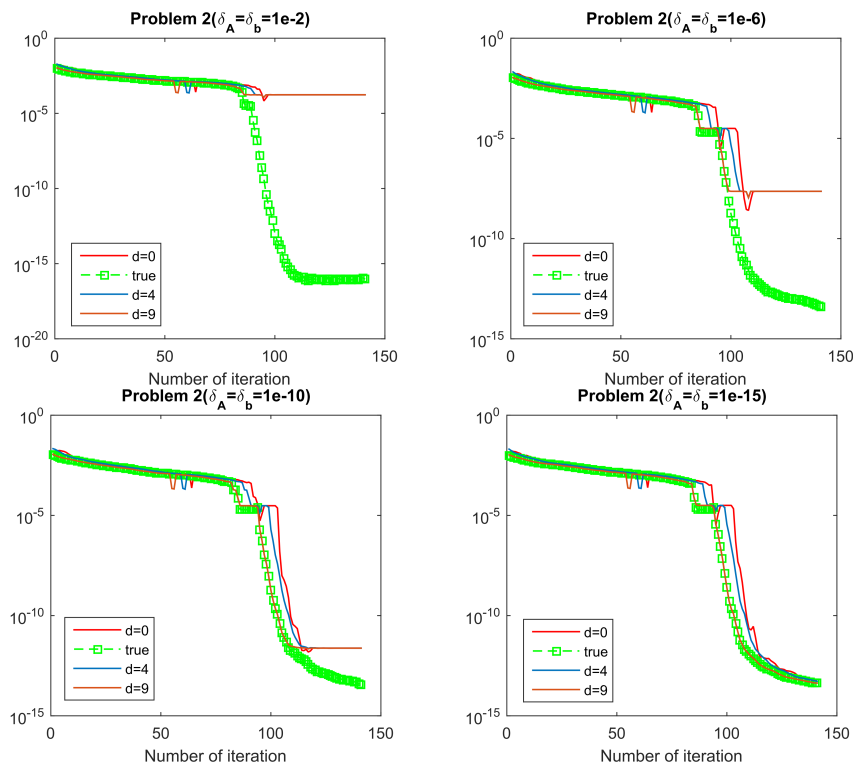


Figure 3. Estimating $\eta(x_k, \gamma, \theta)$ by using $\eta_d(x_k, \gamma, \theta)$ with different d for an ill-conditioned problem.

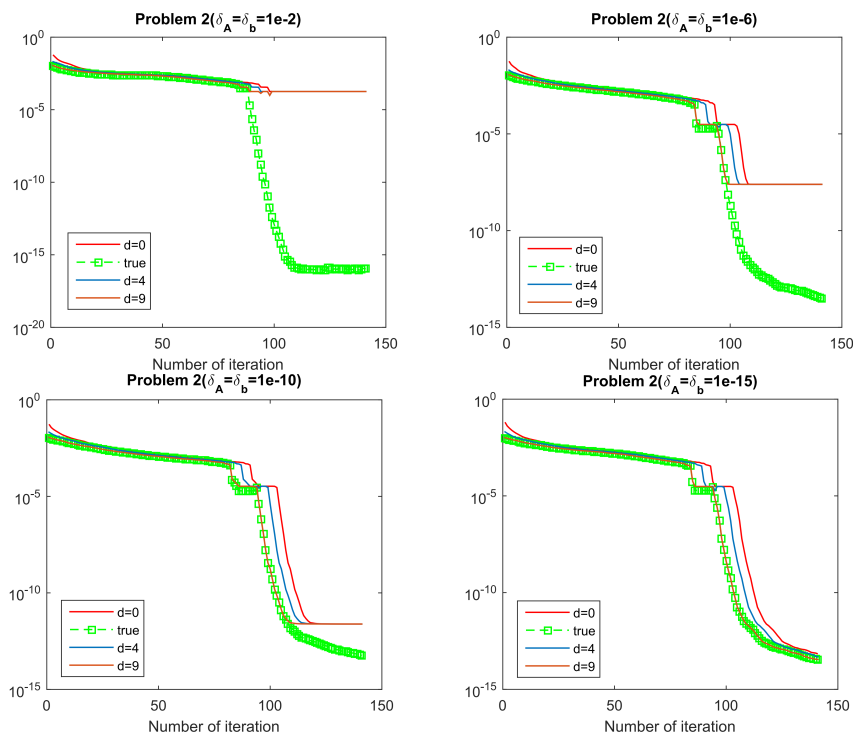


Figure 4. Estimating $\eta(x_k, \gamma, \theta)$ by using $\bar{\eta}_d(x_k, \gamma, \theta)$ with different d for an ill-conditioned problem.

Example 6.3. The problems are taken from Hansen's regularization tools [21]. For example, the case Foxgood is generated by the command $[\bar{A}, \bar{b}, x_{true}] = \text{Shaw}(m)$. Then, noises are added to $\bar{A}$ and $\bar{b}$. Suppose that σ is the relative noise level. We define

$$b = \bar{b} + \sigma \|\bar{b}\| \frac{\xi}{\|\xi\|}, \quad A = \bar{A} + \sigma \|\bar{A}\|_F \frac{Z}{\|Z\|_F},$$

where ξ is a random vector, and Z is a random matrix. Further suppose that

$$\frac{\|b - \bar{b}\|}{\|\bar{b}\|} = \frac{\|A - \bar{A}\|_F}{\|\bar{A}\|_F} = \sigma.$$

Then, we seek the total least squares solution of $Ax \approx b$.

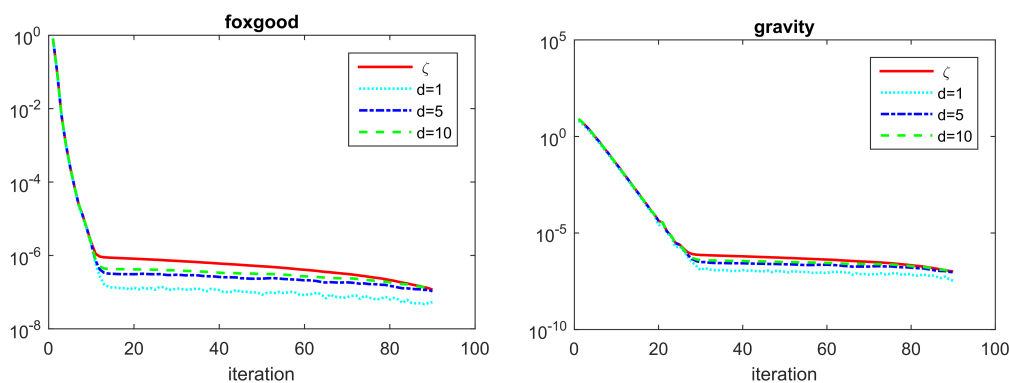


Figure 5. Estimating ζ_k by using $\phi_d(\alpha, \beta)$ in (5.4) with different values of d with $\sigma = 10^{-8}$ for Foxgood (left) and Gravity (right).

Example 6.4. In order to test the effect of our results, the experimental data in [22] are cited, and the coordinate observations are listed in Table 1.

Table 1. Coordinate observations of control points and their corresponding accuracy (unit: m).

	Original coordinates			Target coordinates		
	X_s	Y_s	Z_s	X_t	Y_t	Z_t
1	-2802191.3482	5009064.7657	2772381.1768	-2802088.4182	5009123.1569	2772386.5699
2	-2810175.6515	5016086.1120	2751955.0531	-2810072.7309	5016144.5721	2751960.5007
3	-2820567.2272	5009905.3444	2752470.3858	-2820464.3942	5009963.9917	2752475.9211
4	-2817162.6538	5002454.3520	2769299.1106	-2817059.8351	5002512.9993	2769304.6288
5	-2825775.8182	4995785.4955	2772390.1695	-2825673.0906	4995844.3013	2772395.7629
6	-2821096.7634	4981344.2112	2802 869.4944	-2820994.0605	4981403.0486	2802875.0773
7	-2824710.6654	4984669.2846	2793431.9999	-2824607.9846	4984728.1458	2793437.5997
8	-2827287.5380	4983602.6974	2792671.0817	-2827184.8759	4983661.6016	2792676.7042
9	-2759256.9584	5019419.0725	2796705.2762	-2759153.7202	5019476.8036	2796710.3167
10	-2800063.3339	5001135.2900	2788830.1959	-2799960.4176	5001193.7094	2788835.5802
11	-2841162.8707	4981982.5041	2781464.4489	-2841060.2902	4982041.6183	2781470.1805

Coordinate transformation is the process of transforming spatial data from one coordinate system to another. In essence, it is the process of solving coordinate transformation parameters according to common known points in two coordinate systems. The transformation parameters of the two coordinate systems include three translation parameters, one scale parameter and 3 rotation parameters. When

the rotation angle is small, or the initial value is known, the rotation matrix composed of three rotation angles can be simplified, and the space rectangular coordinates can be transformed into a linear Bursa model [22]. In this example, The condition number of the coefficient matrix corresponding to this problem reaches 10^9 , indicating that the problem is severely ill-conditioned. It is also shown that the upper and lower bounds of the backward error yield better estimates for the problem.

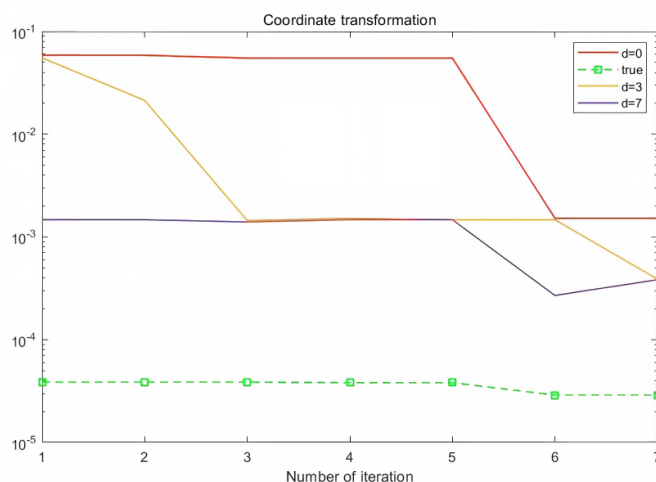


Figure 6. Estimating $\eta(x_k, \gamma, \theta)$ by using $\bar{\eta}_d(x_k, \gamma, \theta)$ with different values of d .

7. Conclusions

In this paper, we show that the backward errors of the STLS problem are equal to those of its core problem. This approach effectively reduces the computational cost of evaluating the backward error, owing to the smaller dimension of the core problem. We have proposed efficient methods to estimate the quantity $\eta(y, \gamma, \theta)$ defined in (2.3), whose direct computation is prohibitively expensive for practical large-scale sparse problems. The backward error $\eta(y, \gamma, \theta)$ can be estimated using the bounds $\underline{\eta}_d(x_k, \gamma, \theta)$ and $\bar{\eta}_d(x_k, \gamma, \theta)$ given in (4.1a) and (4.1b). One may also consider the asymptotic estimates $\underline{\zeta}_d(x_k, \gamma, \theta)$ and $\bar{\zeta}_d(x_k, \gamma, \theta)$ given in (4.8) and (4.9). This strategy is particularly practical for large-scale sparse STLS problems. Finally, we propose practical stopping criteria for the iterative solution of sparse STLS problems. In the Lanczos bidiagonalization TLS algorithm, the estimate for the practical stopping criterion can be computed at virtually no extra cost, albeit at the expense of d additional iterations.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. A. Björck, P. Heggernes, P. Matstoms, Methods for large scale Total Least-Squares problems, *SIAM J. Matrix Anal. Appl.*, **22** (2000), 413–429. <https://doi.org/10.1137/S0895479899355414>
2. C. C. Paige, Z. Strakoš, Scaled total least squares fundamentals, *Numer. Math.*, **91** (2002), 117–146. <https://doi.org/10.1007/s002110100314>
3. C. C. Paige, Z. Strakoš, Unifying least squares, total least squares and data least squares, in *Total Least Squares and Errors-in-Variables Modeling*, (2002), 25–34. https://doi.org/10.1007/978-94-017-3552-0_3
4. B. D. Rao, Unified treatment of LS, TLS, and truncated SVD methods using a weighted TLS framework, in *Proceedings of the Second International Workshop on Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, (1997), 11–20.
5. C. C. Paige, Z. Strakoš, Core problems in linear algebraic systems, *SIAM J. Matrix Anal. Appl.*, **27** (2005), 861–875. <https://doi.org/10.1137/040616991>
6. X. W. Chang, G. H. Golub, C. C. Paige, Towards a backward perturbation analysis for data least squares problems, *SIAM J. Matrix Anal. Appl.*, **30** (2009), 1281–1301. <https://doi.org/10.1137/060668626>
7. X. W. Chang, D. T. Ploquin, Backward perturbation analysis for scaled total least squares problems, *Numer. Linear Algebra Appl.*, **16** (2009), 627–648. <https://doi.org/10.1002/nla.640>
8. E. M. Kasenally, V. Simoncini, Analysis of a minimum perturbation algorithm for nonsymmetric linear systems, *SIAM J. Numer. Anal.*, **34** (1997), 48–66. <https://doi.org/10.1137/S0036142994266844>
9. J. G. Sun, *Perturbations Analysis of Matrix*, Peking, 2001.
10. B. Waldén, R. Karlson, J. G. Sun, Optimal backward perturbation bounds for the linear least squares problem, *Numer. Linear Algebra Appl.*, **2** (1995), 271–286. <https://doi.org/10.1002/nla.1680020308>
11. J. Shan, Y. Wei, Optimal backward error of a total least squares and its randomized algorithms, *SIAM J. Matrix Anal. Appl.*, **3** (2025), 2116–2139. <https://doi.org/10.1137/25M1749657>
12. J. Shan, Y. Wei, Efficient estimate for the optimal backward error of the multidimensional total least squares, *SIAM J. Matrix Anal. Appl.*, **2** (2026), 649–672. <https://doi.org/10.1137/25M1788312>
13. P. Jiránek, D. Tittley-Ploquin, Estimating the backward error in LSQR, *SIAM J. Matrix Anal. Appl.*, **31** (2010), 2055–2074. <https://doi.org/10.1137/090770655>
14. G. H. Golub, W. M. Kahan, Calculating the singular values and pseudo-inverse of a matrix, *J. Soc. Ind. Appl. Math. Ser. B*, **2** (1965), 205–224. <https://doi.org/10.1137/0702016>
15. P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM: Philadelphia, 1998. <https://doi.org/10.1137/1.9780898719697>
16. S. Van Huffel, J. Vandewalle, A. Haegemans, An efficient and reliable algorithm for computing the singular subspace of a matrix, associated with its smallest singular values, *J. Comput. Appl. Math.*, **19** (1987), 313–330. [https://doi.org/10.1016/0377-0427\(87\)90201-9](https://doi.org/10.1016/0377-0427(87)90201-9)
17. J. Baglama, L. Reichel, Augmented implicitly restarted Lanczos bidiagonalization methods, *SIAM J. Sci. Comput.*, **27** (2005), 19–42. <https://doi.org/10.1137/04060593X>

Schematically, the matrix $L_{k,d}$ in (4.8) has the structure

[illegible]

Because of the special structure of $L_{k,d}$, its smallest singular value can be computed using only $O(k + d)$ flops and storing only $O(k + d)$ extra numbers. First, we make QR decomposition of rank one correction matrix for the submatrix of the left upper partition of the matrix above. Then, we apply Givens rotations from the right to eliminate the elements of the block $\omega(I - e_{k+1}e_{k+1}^T)$ one by one. The first k elements are zeroed with the corresponding diagonal of $L_{k,d}$, proceeding upwards from the k th element to the first. To avoid padding, each new nonzero element is immediately zeroed with its neighbor. Below, we illustrate schematically the first two such Givens rotations for the case $k = 3, d = 2$. We use $*$'s to represent nonzero elements and $\bullet$'s for the elements that are modified at the step shown.

The diagram shows a 10x10 grid of points. The points are arranged in a regular pattern. A horizontal line connects a black dot at (4, 3) to a white dot at (6, 3), labeled '1'. A horizontal line connects a black dot at (5, 4) to a white dot at (6, 4), labeled '2'. The origin (0, 0) is marked with a '0'. The grid is enclosed in large parentheses.

The first k elements of the $\omega(I - e_{k+1}e_{k+1}^T)$ block have been once zeroed; a similar process is applied to the last d elements, proceeding down elements $k + 2$ through $k + d + 1$. We illustrate the first two such Givens rotations below:

$$\begin{pmatrix} * & * & * & & 0 & & * \\ & * & * & & 0 & & * \\ & & * & * & & 0 & * \\ & & & * & & 0 & * \\ & & & * & \bullet & & \odot 1 \\ & & & \bullet & & 2 \odot \bullet & * \end{pmatrix}$$

Then, we apply Givens rotations from the right to eliminate the elements of the last column one by one, proceeding downwards from the first element to the $(k+d)$ -th,

$$\begin{pmatrix} * & * & \bullet & & 0 & & \odot \\ & * & \bullet & & 0 & & \bullet \\ & & \bullet & * & & 0 & \bullet \\ & & \bullet & * & & 0 & \bullet \\ & & \bullet & * & * & & 0 & \bullet \\ & & \bullet & & * & & & * & \bullet \end{pmatrix}$$

and we use the last element to turn the last second elements in the last row to zero.

$$\begin{pmatrix} * & * & * & & 0 & & 0 \\ & * & * & & 0 & & 0 \\ & & * & * & & 0 & 0 \\ & & * & * & & 0 & 0 \\ & & * & * & * & & 0 & 0 \\ & & * & & * & & & \odot \bullet \end{pmatrix}$$

With the corresponding zero diagonal element, the bottom right element is merely permuted.

$$\begin{pmatrix} * & * & * & & 0 & & 0 \\ & * & * & & 0 & & 0 \\ & & * & * & & 0 & 0 \\ & & * & * & & 0 & 0 \\ & & * & * & * & & 0 \\ & & * & & * & & 0 \end{pmatrix}$$

After the entire $\omega(I - e_{k+1}e_{k+1}^T)$ block and the last column have been eliminated, Givens rotations can be applied from the left to bring the matrix to bidiagonal form:

$$\begin{pmatrix} * & * & * & & 0 & & 0 \\ & * & * & & 0 & & 0 \\ & & \bullet & \bullet & & 0 & 0 \\ 5 \circ \uparrow & & \bullet & \bullet & & 0 & 0 \\ 3 \circ \uparrow & 4 \circ \uparrow & & \bullet & \bullet & & 0 \\ 2 \circ \uparrow & & 1 \circ \uparrow & \bullet & & & 0 \end{pmatrix}$$



AIMS Press

©2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)