NUMERICAL LINEAR ALGEBRA WITH APPLICATIONS
Numer. Linear Algebra Appl. 2009; **16**:627–648
Published online 5 March 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/nla.640

Backward perturbation analysis for scaled total least-squares problems

X.-W. Chang*, † and D. Titley-Peloquin

School of Comptuer Science, McGill University, 3480 University Street, McConnell Engineering Building, Room 318, Montreal, Canada H3A 2A7

SUMMARY

The scaled total least-squares (STLS) method unifies the ordinary least-squares (OLS), the total least-squares (TLS), and the data least-squares (DLS) methods. In this paper we perform a backward perturbation analysis of the STLS problem. This also unifies the backward perturbation analyses of the OLS, TLS and DLS problems. We derive an expression for an extended minimal backward error of the STLS problem. This is an asymptotically tight lower bound on the true minimal backward error. If the given approximate solution is close enough to the true STLS solution (as is the goal in practice), then the extended minimal backward error is in fact the minimal backward error. Since the extended minimal backward error is expensive to compute directly, we present a lower bound on it as well as an asymptotic estimate for it, both of which can be computed or estimated more efficiently. Our numerical examples suggest that the lower bound gives good order of magnitude approximations, while the asymptotic estimate is an excellent estimate. We show how to use our results to easily obtain the corresponding results for the OLS and DLS problems in the literature. Copyright © 2009 John Wiley & Sons, Ltd.

Received 8 October 2008; Revised 23 January 2009; Accepted 1 February 2009

KEY WORDS: scaled total least squares; backward perturbation analysis

1. INTRODUCTION

Given an approximate solution to a certain problem, backward perturbation analysis involves finding a perturbation in the data of minimal size such that the approximate solution is an exact solution of the perturbed problem. The size of the minimal perturbation is referred to as the minimal backward error. In matrix computations, backward perturbation analyses are useful in two respects.

*Correspondence to: X.-W. Chang, School of Comptuer Science, McGill University, 3480 University Street, McConnell Engineering Building, Room 318, Montreal, Canada H3A 2A7.
†E-mail: chang@cs.mcgill.ca

Contract/grant sponsor: NSERC of Canada; contract/grant number: RGPIN217191-07
Contract/grant sponsor: NSERC of Canada; contract/grant number: PGS-D3

One is to check if a computed solution is a backward stable solution. Sometimes we may not know if an algorithm for solving a problem is numerically stable, but if the relative minimal backward error is of the order of unit round-off, then the computed solution is a backward stable solution and we can be satisfied with it. The other is to use backward perturbation analysis results to design effective stopping criteria for the iterative solution of large sparse problems; see, e.g. [1–4].

There has been a lot of work on the backward perturbation analysis of matrix problems, especially in recent years. Interested readers can find some references on backward perturbation analysis of linear systems (including least-squares problems) in [5].

This paper will give a backward perturbation analysis for scaled total least-squares (STLS) problems. STLS unifies ordinary least squares (OLS), total least squares (TLS) and data least squares (DLS); see Paige and Strakoš [6, 7] and Rao [8]. The backward perturbation analysis of STLS to be given in this paper also unifies the backward perturbation analyses of OLS, TLS and DLS, thus this paper unifies and generalizes the work in [5, 9]. We derive formulas for an 'extended' minimal backward error in Section 3. This extended minimal backward error is at worst a lower bound on the minimal backward error. But we show both in theory (see Section 3) and using numerical tests (see Section 7) that if the given approximate solution is a good enough approximation to the exact solution of the STLS problem, then the extended minimal backward error is the actual minimal backward error. It is time consuming to compute the extended minimal backward error directly, hence in Sections 5 and 6 we derive a lower bound on it and an asymptotic estimate for it, respectively, both of which can be computed or estimated more efficiently.

From these results, we show how to easily obtain the groundbreaking minimal backward error results for OLS problems given by Waldén *et al.* [9], and the extended minimal backward error results for DLS problems given by Chang *et al.* [5]; see Sections 4–6.

We use $I=[e_1,\ldots,e_n]$ to denote the unit matrix. For any matrix $B\in\mathbb{R}^{m\times n}$, its 2-norm and F-norm are denoted by $\|B\|_2$ and $\|B\|_F$, respectively, its Moore–Penrose generalized inverse is denoted by $B^\dagger$, its smallest singular value (the pth largest singular value with $p=\min\{m,n\}$) by $\sigma_{\min}(B)$, its smallest eigenvalue is denoted by $\lambda_{\min}$ (when B is symmetric) and its condition number in the 2-norm by $\kappa_2(B)$. For any matrix $B=[b_1,\ldots,b_n]$, define $\mathrm{vec}(B)=[b_1^{\mathrm{T}},\ldots,b_n^{\mathrm{T}}]^{\mathrm{T}}$. For any vector $v\in\mathbb{R}^n$, its 2-norm is denoted by $\|v\|$ and its Moore–Penrose generalized inverse is

$$v^\dagger\equiv\begin{cases}0 & \text{if } v=0\\ v^{\mathrm{T}}/\|v\|^2 & \text{if } v\neq 0\end{cases}$$

2. SCALED TOTAL LEAST-SQUARES PROBLEMS

For given $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$ and $\gamma>0$, the scaled total least-squares (STLS) problem with data $[A,b]$ can be formulated as (see [8])

$$\text{STLS distance}\equiv\sigma_S\equiv\min_{E,f,x}\|[E,f\gamma]\|_F\quad\text{subject to }(A+E)x=b+f \tag{1}$$

A theoretically equivalent but different formulation can be found in [6]. The STLS problem (1) may not have a solution, but it does have a unique solution if the following condition holds (see [6], (1.11)):

$$\mathrm{rank}(A)=n\quad\text{and}\quad b\not\perp\mathscr{U}_{\min} \tag{2}$$

DOI: 10.1002/nla

where $\mathscr{U}_{\min}$ is the left singular vector subspace of A corresponding to its minimal singular value $\sigma_{\min}(A)$. From now on, we assume that (2) holds.

It was shown in [10] that σ_S^2 in (1) is the global minimum of the function

$$\sigma^2(x) \equiv \frac{\|b-Ax\|^2}{\gamma^{-2}+\|x\|^2} \tag{3}$$

It can be proven (see [11, Theorem 2.7]; [6, Section 6]) that when (2) holds, $\hat{x}$ solves (1) if and only if it satisfies

$$A^{\mathrm{T}}(b-A\hat{x}) = -\frac{\|b-A\hat{x}\|^2}{\gamma^{-2}+\|\hat{x}\|^2}\hat{x} \tag{4}$$

$$\frac{\|b-A\hat{x}\|^2}{\gamma^{-2}+\|\hat{x}\|^2} < \sigma_{\min}^2(A) \tag{5}$$

Note that (4) can be obtained by setting the derivative of $\sigma^2(x)$ in (3) equal to zero. All stationary points of $\sigma^2(x)$ therefore satisfy (4). The global minimum must also satisfy (5). It is easy to observe from (4) and (5) (see also [6, Section 6]) that when $\gamma=1$, $\hat{x}$ becomes the TLS solution, when $\gamma\to 0$, $\hat{x}$ converges to the OLS solution (note that in this case the left-hand side of (5) becomes zero and (5) holds automatically), and when $\gamma\to\infty$, $\hat{x}$ converges to the DLS solution.

3. BACKWARD ERROR ANALYSIS

Suppose we have obtained a nonzero approximation $y\in\mathbb{R}^n$ to the solution vector $\hat{x}$ of (1). In order to find the closest STLS problem whose solution is actually y, we see from (3) that we need to solve a backward error problem of the form

$$\min_{\Delta A,\Delta b} \|[\Delta A, \Delta b\,\theta]\|_F \quad \text{subject to } y = \arg\min_x \frac{\|b+\Delta b-(A+\Delta A)x\|^2}{\gamma^{-2}+\|x\|^2} \tag{6}$$

Here, the chosen scalar $\theta>0$ (different, sometimes, from γ in (1)) allows a different emphasis on each data error. The above is the objective function that we use here to define 'closest'.

In order to solve (6), we first need to characterize the set of $[\Delta A, \Delta b]$ satisfying the equality in (6). It follows from (4) and (5) that 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:

$$\begin{aligned}\mathscr{C}_{A,b}(\gamma) \equiv \Bigg\{ &[\Delta A, \Delta b] : (A+\Delta A)^{\mathrm{T}}[b+\Delta b-(A+\Delta A)y] \\ &= \frac{\|b+\Delta b-(A+\Delta A)y\|^2}{\gamma^{-2}+\|y\|^2}y,\ \frac{\|b+\Delta b-(A+\Delta A)y\|^2}{\gamma^{-2}+\|y\|^2} < \sigma_{\min}^2(A+\Delta A)\Bigg\}\end{aligned} \tag{7}$$

A natural idea to solve (6) is to find an explicit expression for the set $\mathscr{C}_{A,b}(\gamma)$ and then minimize $\|[\Delta A, \Delta b\,\theta]\|_F$ over this set. Unfortunately, the inequality in (7) makes it difficult to derive a general expression for $[\Delta A, \Delta b]\in\mathscr{C}_{A,b}(\gamma)$, hence we initially ignore it and consider a larger set

$$\mathscr{C}_{A,b}^{+}(\gamma) \equiv \left\{ [\Delta A, \Delta b] : (A+\Delta A)^{\mathrm{T}}[b+\Delta b-(A+\Delta A)y] = -\frac{\|b+\Delta b-(A+\Delta A)y\|^2}{\gamma^{-2}+\|y\|^2}y \right\} \tag{8}$$

DOI: 10.1002/nla

The following result from Theorem 3.2 of [12] gives an explicit relation between ΔA and Δb for any $[\Delta A, \Delta b] \in \mathscr{C}^{+}_{A,b}(\gamma)$.

Lemma 3.1
$[\Delta A, \Delta b] \in \mathscr{C}^{+}_{A,b}(\gamma)$ if and only if

$$A + \Delta A = -\gamma^2 w w^{\dagger}(b + \Delta b) y^{\mathrm{T}} + (I - w w^{\dagger})[(b + \Delta b) y^{\dagger} + Z(I - y y^{\dagger})] \tag{9}$$

for $w \in \mathbb{R}^m$ and $Z \in \mathbb{R}^{m \times n}$.

Based on Lemma 3.1, we now solve the extended minimal backward error problem

$$\mu(y, \gamma, \theta) \equiv \min_{[\Delta A, \Delta b] \in \mathscr{C}^{+}_{A,b}(\gamma)} \|[\Delta A, \Delta b\, \theta]\|_F \tag{10}$$

Theorem 3.1
Suppose we are given $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, nonzero approximate STLS solution $y \in \mathbb{R}^n$, $\gamma > 0$ and $\theta > 0$; and suppose that (2) holds. Let $r \equiv b - Ay$ and

$$N \equiv \left[A(I - y y^{\dagger}), \frac{\theta \|r\|}{\sqrt{1 + \theta^2 \|y\|^2}} (I - r r^{\dagger}), \frac{\theta}{\sqrt{\theta^2 \|y\|^2 + \gamma^4 \|y\|^4}} (Ay + \gamma^2 \|y\|^2 b) \right] \tag{11}$$

$$M \equiv M(y, \gamma, \theta) \equiv A(I - y y^{\dagger}) A^{\mathrm{T}} - \frac{\theta^2}{1 + \theta^2 \|y\|^2} r r^{\mathrm{T}} + \frac{\theta^2}{\theta^2 \|y\|^2 + \gamma^4 \|y\|^4} (Ay + \gamma^2 \|y\|^2 b)$$

$$\times (Ay + \gamma^2 \|y\|^2 b)^{\mathrm{T}} = N N^{\mathrm{T}} - \frac{\theta^2 \|r\|^2}{1 + \theta^2 \|y\|^2} I \tag{12}$$

Then M has at most one negative eigenvalue, and

$$\mu^2(y, \gamma, \theta) = \begin{cases} \dfrac{\theta^2 \|r\|^2}{1 + \theta^2 \|y\|^2} & \text{if } \lambda_{\min}(M) \geqslant 0 \\ \dfrac{\theta^2 \|r\|^2}{1 + \theta^2 \|y\|^2} + \lambda_{\min}(M) = \sigma^2_{\min}(N) & \text{if } \lambda_{\min}(M) < 0 \end{cases} \tag{13}$$

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

$$\widehat{\Delta A} = \begin{cases} \dfrac{\theta^2}{1 + \theta^2 \|y\|^2} r y^{\mathrm{T}} & \text{if } \lambda_{\min}(M) \geqslant 0 \\ \dfrac{\theta^2}{1 + \theta^2 \|y\|^2} r y^{\mathrm{T}} - w_* w_*^{\mathrm{T}} \left[A + \dfrac{\gamma^2 (\theta^2 - \gamma^2)^2}{\theta} + \gamma^4 \|y\|^2 b y^{\mathrm{T}} \right. & \\ \quad \left. + \dfrac{\gamma^4 + 2\gamma^4 \theta^2 \|y\|^2 + \theta^4}{(1 + \theta^2 \|y\|^2)(\theta^2 + \gamma^4 \|y\|^2)} r y^{\mathrm{T}} \right] & \text{if } \lambda_{\min}(M) < 0 \end{cases} \tag{14}$$

DOI: 10.1002/nla

$$\widehat{\Delta b}=\begin{cases}-\dfrac{1}{1+\theta^2\|y\|^2}r & \text{if } \lambda_{\min}(M)\geqslant 0\\ -\dfrac{1}{1+\theta^2\|y\|^2}r-w_*w_*^{\mathrm{T}}\left[\dfrac{\gamma^2(1+\gamma^2\|y\|^2)}{\theta^2+\gamma^4\|y\|^2}b\right. & \\ \quad\left.-\dfrac{(\gamma^2+\theta^2)(1+\gamma^2\|y\|^2)}{(1+\theta^2\|y\|^2)(\theta^2+\gamma^4\|y\|^2)}r\right] & \text{if } \lambda_{\min}(M)<0\end{cases} \tag{15}$$

where w_* is the unit eigenvector of M corresponding to $\lambda_{\min}(M)$, or equivalently the unit left singular vector of N corresponding to $\sigma_{\min}(N)$.

Proof

On the right-hand side of the second equality of (12), both the first and the third terms are nonnegative definite, while the second term is a rank one matrix. By [13, Theorem 4.3.4(b)] with $k=1$, M has at most one negative eigenvalue.

From Lemma 3.1, we see that ΔA and Δb are functions of w and Z. Therefore, to solve (10), we need to find the optimal w and Z. We discuss two cases separately.

Case 1: The optimal $w=0$. The proof is almost the same as the corresponding part in the proof of Theorem 2.2 in [5]. But for readers' convenience, we give it here. In this case, from (9) we have

$$\Delta A=(b+\Delta b)y^{\dagger}+Z(I-yy^{\dagger})-A \tag{16}$$

Given a nonzero y, we can find $Y_2\in\mathbb{R}^{n\times(n-1)}$ such that $Y=[y/\|y\|, Y_2]$ is orthogonal. Thus

$$\Delta AY=[(b+\Delta b)/\|y\|, 0]+Z[0, Y_2]-[Ay/\|y\|, AY_2]=[(r+\Delta b)/\|y\|, (Z-A)Y_2]$$

Therefore, we have

$$\begin{aligned}\|[\Delta A, \Delta b\,\theta]\|_F^2&=\|r+\Delta b\|^2/\|y\|^2+\|(Z-A)Y_2\|_F^2+\theta^2\|\Delta b\|^2\\&=\frac{1}{\|y\|^2}\left\|\begin{bmatrix}I\\ \theta\|y\|I\end{bmatrix}\Delta b+\begin{bmatrix}r\\0\end{bmatrix}\right\|^2+\|(Z-A)Y_2\|_F^2\end{aligned}$$

It follows that the optimal $\widehat{\Delta b}$ and $\widehat{Z}$ satisfy

$$\widehat{\Delta b}=-\begin{bmatrix}I\\ \theta\|y\|I\end{bmatrix}^{\dagger}\begin{bmatrix}r\\0\end{bmatrix}=-\frac{1}{1+\theta^2\|y\|^2}r,\quad \widehat{Z}=A \tag{17}$$

Then from (16) the optimal $\widehat{\Delta A}$ satisfies

$$\widehat{\Delta A}=\left(b-\frac{1}{1+\theta^2\|y\|}\right)ry^{\dagger}-Ayy^{\dagger}=\frac{\theta^2}{1+\theta^2\|y\|^2}ry^{\mathrm{T}} \tag{18}$$

leading to

$$\|[\widehat{\Delta A}, \widehat{\Delta b}\,\theta]\|_F^2=\frac{\theta^2\|r\|^2}{1+\theta^2\|y\|^2} \tag{19}$$

Case 2: The optimal $w\neq 0$. Let Y be as in Case 1. Since $ww^{\dagger}$ is independent of the length of w, we can assume that $\|w\|=1$ in (9). Construct $W=[w, W_2]\in\mathbb{R}^{m\times m}$ such that it is orthogonal.

DOI: 10.1002/nla

Then we have from (9)

$$W^{\mathrm{T}}\Delta A Y=\begin{bmatrix} -\gamma^2 w^{\mathrm{T}}(b+\Delta b)\|y\|-w^{\mathrm{T}}Ay/\|y\| & -w^{\mathrm{T}}AY_2 \\ W_2^{\mathrm{T}}(b+\Delta b)/\|y\|-W_2^{\mathrm{T}}Ay/\|y\| & W_2^{\mathrm{T}}ZY_2-W_2^{\mathrm{T}}AY_2 \end{bmatrix}$$

Thus, it follows that

$$\begin{aligned}\|[\Delta A,\Delta b\,\theta]\|_F^2 &= [\gamma^2\|y\|w^{\mathrm{T}}(b+\Delta b)+w^{\mathrm{T}}Ay/\|y\|]^2+\|w^{\mathrm{T}}AY_2\|_F^2 \\ &\quad +\|W_2^{\mathrm{T}}(r+\Delta b)\|^2/\|y\|^2+\|W_2^{\mathrm{T}}(Z-A)Y_2\|_F^2+\theta^2\|\Delta b\|^2 \\ &= [\gamma^2\|y\|w^{\mathrm{T}}(b+\Delta b)+w^{\mathrm{T}}Ay/\|y\|]^2+\|w^{\mathrm{T}}A(I-yy^{\dagger})\|_F^2 \\ &\quad +\|(I-ww^{\mathrm{T}})(r+\Delta b)\|^2/\|y\|^2+\|W_2^{\mathrm{T}}(Z-A)Y_2\|_F^2+\theta^2\|\Delta b\|^2 \\ &= \frac{1}{\|y\|^2}\left\|\begin{bmatrix}\gamma^2\|y\|^2w^{\mathrm{T}}\\ I-ww^{\mathrm{T}}\\ \theta\|y\|I\end{bmatrix}\Delta b+\begin{bmatrix}\gamma^2\|y\|^2w^{\mathrm{T}}b+w^{\mathrm{T}}Ay\\ (I-ww^{\mathrm{T}})r\\ 0\end{bmatrix}\right\|^2 \\ &\quad +\|w^{\mathrm{T}}A(I-yy^{\dagger})\|_F^2+\|W_2^{\mathrm{T}}(Z-A)Y_2\|_F^2 \end{aligned} \tag{20}$$

Obviously $\widehat{Z}=A$ is optimal, and if w is fixed, the optimal $\widehat{\Delta b}$ satisfies

$$\begin{aligned}\widehat{\Delta b} &= -\left(\begin{bmatrix}\gamma^2\|y\|^2w^{\mathrm{T}}\\ I-ww^{\mathrm{T}}\\ \theta\|y\|I\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\gamma^2\|y\|^2w^{\mathrm{T}}\\ I-ww^{\mathrm{T}}\\ \theta\|y\|I\end{bmatrix}\right)^{-1}\begin{bmatrix}\gamma^2\|y\|^2w^{\mathrm{T}}\\ I-ww^{\mathrm{T}}\\ \theta\|y\|I\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\gamma^2\|y\|^2w^{\mathrm{T}}b+w^{\mathrm{T}}Ay\\ (I-ww^{\mathrm{T}})r\\ 0\end{bmatrix} \\ &= -\frac{1}{\theta^2+\gamma^4\|y\|^2}ww^{\mathrm{T}}(\gamma^2Ay+\gamma^4\|y\|^2b)-\frac{1}{1+\theta^2\|y\|^2}(I-ww^{\mathrm{T}})r \end{aligned} \tag{21}$$

where we used the Sherman–Morrisson–Woodbury formula to simplify the inverse. Thus, with the above $\widehat{Z}$ and $\widehat{\Delta b}$, using $w^{\mathrm{T}}w=1$ we have from (20) that

$$\begin{aligned}\|[\Delta A,\widehat{\Delta b}\,\theta]\|_F^2 &= \frac{\theta^2}{\theta^2\|y\|^2+\gamma^4\|y\|^4}w^{\mathrm{T}}(Ay+\gamma^2\|y\|^2b)(Ay+\gamma^2\|y\|^2b)^{\mathrm{T}}w \\ &\quad +\frac{\theta^2}{1+\theta^2\|y\|^2}r^{\mathrm{T}}(I-ww^{\mathrm{T}})r+w^{\mathrm{T}}A(I-yy^{\dagger})A^{\mathrm{T}}w \\ &= w^{\mathrm{T}}\left[A(I-yy^{\dagger})A^{\mathrm{T}}+\frac{\theta^2\|r\|^2}{1+\theta^2\|y\|^2}(I-rr^{\dagger})\right. \\ &\quad \left.+\frac{\theta^2}{\theta^2\|y\|^2+\gamma^4\|y\|^4}(Ay+\gamma^2\|y\|^2b)(Ay+\gamma^2\|y\|^2b)^{\mathrm{T}}\right]w \\ &= w^{\mathrm{T}}NN^{\mathrm{T}}w=\frac{\theta^2\|r\|^2}{1+\theta^2\|y\|^2}+w^{\mathrm{T}}Mw \end{aligned} \tag{22}$$

DOI: 10.1002/nla

Then the minimal value is reached when w is equal to w_*, the unit left singular vector of N corresponding to its smallest singular value $\sigma_{\min}(N)$, or equivalently, the unit eigenvector of M corresponding to its smallest eigenvalue.

If $\lambda_{\min}(M)\geqslant 0$, from (19) and (22), we can see that the minimum value of $\|[\Delta A, \widehat{\Delta b}\,\theta]\|_F$ is reached in Case 1 and the top equalities in (13), (14) and (15) follow from (19), (18) and (17), respectively. If $\lambda_{\min}(M)<0$, from (19) and (22), we can see that the minimum value of $\|[\Delta A, \widehat{\Delta b}\,\theta]\|_F$ is reached in Case 2 and the bottom equality in (13) follows from (22). In this case, the bottom equality in (15) can be obtained from (21) by some simple algebraic operations and the bottom equality in (14) can easily be proved from (9) with the optimal w, Δb and Z found in Case 2. □

Corollary 3.1
With the notation and conditions of Theorem 3.1 and the STLS solution $\hat{x}$ of (4) and (5), define $\widehat{M} \equiv M(\hat{x}, \gamma, \theta)$ (see (12)) and $\hat{r} \equiv b - A\hat{x}$. Then

$$\lim_{y \to \hat{x}} \mu(y, \gamma, \theta) = \mu(\hat{x}, \gamma, \theta) = 0 \tag{23}$$

Proof
First, we consider the case that $\hat{r} \neq 0$. Multiplying both sides of (4) by $\hat{x}$ from the left, we obtain

$$\hat{x}^{\mathrm{T}} A^{\mathrm{T}} \hat{r} = -\frac{\|\hat{x}\|^2 (b - A\hat{x})^{\mathrm{T}} \hat{r}}{\gamma^{-2} + \|\hat{x}\|^2} = \frac{\|\hat{x}\|^2}{\gamma^{-2} + \|\hat{x}\|^2} \hat{x}^{\mathrm{T}} A^{\mathrm{T}} \hat{r} - \frac{\|\hat{x}\|^2}{\gamma^{-2} + \|\hat{x}\|^2} b^{\mathrm{T}} \hat{r}$$

leading to $\hat{x}^{\mathrm{T}} A^{\mathrm{T}} \hat{r} + \gamma^2 \hat{x}^{\mathrm{T}} \hat{x} b^{\mathrm{T}} \hat{r} = 0$. Multiplying both sides of (4) by $I - \hat{x}\hat{x}^{\dagger}$ from the left, we immediately obtain $(I - \hat{x}\hat{x}^{\dagger}) A^{\mathrm{T}} \hat{r} = 0$. Therefore, from (12),

$$\begin{aligned} \widehat{M}\hat{r} &= A(I - \hat{x}\hat{x}^{\dagger}) A^{\mathrm{T}} \hat{r} - \frac{\theta^2 \|\hat{r}\|^2}{1 + \theta^2 \|\hat{x}\|^2} \hat{r} + \frac{\theta^2}{\theta^2 \|\hat{x}\|^2 + \gamma^4 \|\hat{x}\|^4} (A\hat{x} + \gamma^2 \|\hat{x}\|^2 b)(\hat{x}^{\mathrm{T}} A^{\mathrm{T}} + \gamma^2 \|\hat{x}\|^2 b^{\mathrm{T}}) \hat{r} \\ &= \frac{-\theta^2 \|\hat{r}\|^2}{1 + \theta^2 \|\hat{x}\|^2} \hat{r} \equiv \hat{\lambda} \hat{r} \end{aligned}$$

But by Theorem 3.1, $\widehat{M}$ has at most one negative eigenvalue. Thus, $\lambda_{\min}(\widehat{M}) = \hat{\lambda} < 0$ and from the bottom of (13) $\mu(\hat{x}, \gamma, \theta) = 0$. When y is close enough to $\hat{x}$, $\lambda_{\min}(M) < 0$. Therefore,

$$\lim_{y \to \hat{x}} \mu(y, \gamma, \theta) = \lim_{y \to \hat{x}} \left(\frac{\theta^2 \|r\|^2}{1 + \theta^2 \|y\|^2} + \lambda_{\min}(M) \right)^{1/2} = \left(\frac{\theta^2 \|\hat{r}\|^2}{1 + \theta^2 \|\hat{x}\|^2} + \lambda_{\min}(\widehat{M}) \right)^{1/2} = 0$$

Now we consider the case that $\hat{r} = 0$. Since $\mu(\hat{x}, \gamma, \theta) \geqslant 0$, from (13) we must have $\lambda_{\min}(\widehat{M}) \geqslant 0$. Thus from (13)

$$\begin{aligned} \lim_{y \to \hat{x}} \mu(y, \gamma, \theta) &= \lim_{y \to \hat{x}} \min \left\{ \left(\frac{\theta^2 \|r\|^2}{1 + \theta^2 \|y\|^2} \right)^{1/2}, \left(\frac{\theta^2 \|r\|^2}{1 + \theta^2 \|y\|^2} + \lambda_{\min}(M) \right)^{1/2} \right\} \\ &= \min \left\{ \left(\frac{\theta^2 \|\hat{r}\|^2}{1 + \theta^2 \|\hat{x}\|^2} \right)^{1/2}, \left(\frac{\theta^2 \|\hat{r}\|^2}{1 + \theta^2 \|\hat{x}\|^2} + \lambda_{\min}(\widehat{M}) \right)^{1/2} \right\} \\ &= \mu(\hat{x}, \gamma, \theta) = 0 \end{aligned}$$

□

DOI: 10.1002/nla

Since $\mathscr{C}_{A,b}(\gamma) \subseteq \mathscr{C}^{+}_{A,b}(\gamma)$,

$$\mu(y,\gamma,\theta) \equiv \min_{[\Delta A,\Delta b]\in\mathscr{C}^{+}_{A,b}(\gamma)} \|[\Delta A, \Delta b\,\theta]\|_F \leqslant \min_{[\Delta A,\Delta b]\in\mathscr{C}_{A,b}(\gamma)} \|[\Delta A, \Delta b\,\theta]\|_F$$

The following theorem, which is analogous to Theorem 2.8 of [5] states that if y is close enough to the true solution $\hat{x}$, as is the goal in practice, then $\mu(y,\gamma,\theta)$ is in fact the minimal backward error.

Theorem 3.2
With the notation and conditions of Theorem 3.1 and the STLS solution $\hat{x}$ of (4) and (5), there exists $\varepsilon>0$ such that for all y satisfying $\|y-\hat{x}\|\leqslant\varepsilon$, $\mu(y,\gamma,\theta)$ is the true minimal backward error.

Proof
The proof is similar to that for Theorem 2.8 of [5]. For any given y, Theorem 3.1 shows that $\widehat{\Delta A}$ satisfying (14) and $\widehat{\Delta b}$ satisfying (15) are the minimizers of (10). Notice that when $y\to\hat{x}$ we have from Corollary 3.1 that $\mu(y,\gamma,\theta)\to 0$, in other words $\widehat{\Delta A}\to 0$ and $\widehat{\Delta b}\to 0$. Thus

$$\lim_{y\to\hat{x}}\left(\frac{\|b+\widehat{\Delta b}-(A+\widehat{\Delta A})y\|^2}{\gamma^{-2}+\|y\|^2}-\sigma^2_{\min}(A+\widehat{\Delta A})\right)=\frac{\|b-A\hat{x}\|^2}{\gamma^{-2}+\|\hat{x}\|^2}-\sigma^2_{\min}(A)$$

Since

$$\frac{\|b-A\hat{x}\|^2}{\gamma^{-2}+\|\hat{x}\|^2}-\sigma^2_{\min}(A)<0$$

see (5), there must exist $\varepsilon>0$ such that when $\|y-\hat{x}\|<\varepsilon$,

$$\frac{\|b+\widehat{\Delta b}-(A+\widehat{\Delta A})y\|^2}{\gamma^{-2}+\|y\|^2}-\sigma^2_{\min}(A+\widehat{\Delta A})<0$$

Therefore, when $\|y-\hat{x}\|<\varepsilon$, $[\widehat{\Delta A},\widehat{\Delta b}]\in\mathscr{C}_{A,b}(\gamma)$ and thus $\mu(y,\gamma,\theta)=\min_{[\Delta A,\Delta b]\in\mathcal{C}_{A,b}(\gamma)}\|[\Delta A,\Delta b\theta]\|_F$, hence $\mu(y,\gamma,\theta)$ is in fact the true minimal backward error. □

We will give numerical examples in Section 7 showing that when y is the reasonable approximation to the exact solution of the STLS problem (1), $\widehat{\Delta A}$ and $\widehat{\Delta b}$ usually satisfy the inequality in (7). In such cases, $\mu(y,\gamma,\theta)$ is the actual minimal backward error.

4. MINIMAL BACKWARD ERRORS FOR OLS AND DLS PROBLEMS

The OLS minimal backward error problem and the DLS extended minimal backward error problem (see [5]) are, respectively,

$$\mu_{\mathrm{OLS}}(y,\theta) \equiv \min_{[\Delta A,\Delta b]\in\mathscr{C}^{+}_{A,b}(0)} \|[\Delta A, \Delta b\,\theta]\|_F$$

$$\mu_{\mathrm{DLS}}(y,\theta) \equiv \min_{[\Delta A,\Delta b]\in\mathscr{C}^{+}_{A,b}(\infty)} \|[\Delta A, \Delta b\,\theta]\|_F$$

DOI: 10.1002/nla

where we have used $\mathscr{C}^+_{A,b}(\infty)$ to denote

$$\lim_{\gamma\to\infty}\mathscr{C}^+_{A,b}(\gamma)$$

From (13) we see that $\mu(y,\gamma,\theta)$ is a continuous function of γ. Then from the definition of $\mu(y,\gamma,\theta)$ in (10), we have

$$\mu_{\mathrm{OLS}}(y,\theta)=\lim_{\gamma\to 0}\mu(y,\gamma,\theta),\quad \mu_{\mathrm{DLS}}(y,\theta)=\lim_{\gamma\to\infty}\mu(y,\gamma,\theta) \tag{24}$$

In this section, we show that using Theorem 3.1 we can easily obtain the backward error results for the OLS problem obtained in [9] and the backward error results for the DLS problem obtained in [5].

4.1. Minimal backward error for OLS

From (12) we have

$$M_{\mathrm{OLS}}\equiv\lim_{\gamma\to 0}M(y,\gamma,\theta)=AA^{\mathrm{T}}-\frac{\theta^2}{1+\theta^2\|y\|^2}rr^{\mathrm{T}}$$

Then by the continuity of eigenvalues,

$$\lim_{\gamma\to 0}\lambda_{\min}(M(y,\gamma,\theta))=\lambda_{\min}\left(\lim_{\gamma\to 0}M(y,\gamma,\theta)\right)=\lambda_{\min}(M_{\mathrm{OLS}})$$

Then from (24) and (13), we obtain

$$\mu^2_{\mathrm{OLS}}(y,\theta)=\lim_{\gamma\to 0}\mu^2(y,\gamma,\theta)=\begin{cases}\dfrac{\theta^2\|r\|^2}{1+\theta^2\|y\|^2} & \text{if } \lambda_{\min}(M_{\mathrm{OLS}})\geqslant 0\\[2ex] \dfrac{\theta^2\|r\|^2}{1+\theta^2\|y\|^2}+\lambda_{\min}(M_{\mathrm{OLS}}) & \text{if } \lambda_{\min}(M_{\mathrm{OLS}})<0\end{cases}$$

This is identical to the formula for the OLS minimal backward error, with perturbations in both A and b, derived in [9].

Furthermore, the optimal perturbations $\widehat{\Delta A}$ and $\widehat{\Delta b}$ in (14) and (15) become

$$\widehat{\Delta A}_{\mathrm{OLS}}\equiv\lim_{\gamma\to 0}\widehat{\Delta A}=\begin{cases}\dfrac{\theta^2}{1+\theta^2\|y\|^2}ry^{\mathrm{T}} & \text{if } \lambda_{\min}(M_{\mathrm{OLS}})\geqslant 0\\[2ex] \dfrac{\theta^2}{1+\theta^2\|y\|^2}ry^{\mathrm{T}} & \\ \quad -w_{\mathrm{OLS}}w_{\mathrm{OLS}}^{\mathrm{T}}\left[A+\dfrac{\theta^2}{1+\theta^2\|y\|^2}ry^{\mathrm{T}}\right] & \text{if } \lambda_{\min}(M_{\mathrm{OLS}})<0\end{cases}$$

$$\widehat{\Delta b}_{\mathrm{OLS}}\equiv\lim_{\gamma\to 0}\widehat{\Delta b}=\begin{cases}-\dfrac{1}{1+\theta^2\|y\|^2}r & \text{if } \lambda_{\min}(M_{\mathrm{OLS}})\geqslant 0\\[2ex] -\dfrac{1}{1+\theta^2\|y\|^2}(I-w_{\mathrm{OLS}}w_{\mathrm{OLS}}^{\mathrm{T}})r & \text{if } \lambda_{\min}(M_{\mathrm{OLS}})<0\end{cases}$$

DOI: 10.1002/nla

where w_{OLS} is the unit eigenvector of M_{OLS} corresponding to $\lambda_{\min}(M_{\mathrm{OLS}})$. The formulas for $\widehat{\Delta A}_{\mathrm{OLS}}$ and $\widehat{\Delta b}_{\mathrm{OLS}}$ were also derived in [9].

4.2. Extended minimal backward error for DLS

From (12) we have

$$M_{\mathrm{DLS}} \equiv \lim_{\gamma\to\infty} M(y,\gamma,\theta) = A(I-yy^{\dagger})A^{\mathrm{T}} - \frac{\theta^2}{1+\theta^2\|y\|^2} rr^{\mathrm{T}} + \theta^2 bb^{\mathrm{T}}$$

Once again using the continuity of eigenvalues,

$$\lim_{\gamma\to\infty} \lambda_{\min}(M(y,\gamma,\theta)) = \lambda_{\min}\left(\lim_{\gamma\to\infty} M(y,\gamma,\theta)\right) = \lambda_{\min}(M_{\mathrm{DLS}})$$

Then from (24) and (13), we obtain

$$\mu^2_{\mathrm{DLS}}(y,\theta) \equiv \lim_{\gamma\to\infty} \mu^2(y,\gamma,\theta) = \begin{cases} \dfrac{\theta^2\|r\|^2}{1+\theta^2\|y\|^2}, & \lambda_{\min}(M_{\mathrm{DLS}}) \geqslant 0 \\ \dfrac{\theta^2\|r\|^2}{1+\theta^2\|y\|^2} + \lambda_{\min}(M_{\mathrm{DLS}}), & \lambda_{\min}(M_{\mathrm{DLS}}) < 0 \end{cases}$$

This is the formula for the DLS extended minimal backward error, with perturbations in both A and b, derived in [5].

It is also straightforward to verify that the optimal perturbations $\widehat{\Delta A}$ and $\widehat{\Delta b}$ in (14) and (15) become

$$\widehat{\Delta A}_{\mathrm{DLS}} \equiv \lim_{\gamma\to\infty} \widehat{\Delta A} = \begin{cases} \dfrac{\theta^2}{1+\theta^2\|y\|^2} ry^{\mathrm{T}} & \text{if } \lambda_{\min}(M_{\mathrm{DLS}}) \geqslant 0 \\ \dfrac{\theta^2}{1+\theta^2\|y\|^2} ry^{\mathrm{T}} & \\ \quad - w_{\mathrm{DLS}} w_{\mathrm{DLS}}^{\mathrm{T}} \left[A - by^{\dagger} + \dfrac{1+2\theta^2\|y\|^2}{1+\theta^2\|y\|^2} ry^{\dagger} \right] & \text{if } \lambda_{\min}(M_{\mathrm{DLS}}) < 0 \end{cases}$$

$$\widehat{\Delta b}_{\mathrm{DLS}} \equiv \lim_{\gamma\to\infty} \widehat{\Delta b} = \begin{cases} -\dfrac{1}{1+\theta^2\|y\|^2} r & \text{if } \lambda_{\min}(M_{\mathrm{DLS}}) \geqslant 0 \\ -\dfrac{1}{1+\theta^2\|y\|^2} (I - w_{\mathrm{DLS}} w_{\mathrm{DLS}}^{\mathrm{T}}) r - w_{\mathrm{DLS}} w_{\mathrm{DLS}}^{\mathrm{T}} b & \text{if } \lambda_{\min}(M_{\mathrm{DLS}}) < 0 \end{cases}$$

where w_{DLS} is the unit eigenvector of the matrix M_{DLS} corresponding to its smallest eigenvalue. These formulas were also given in [5].

DOI: 10.1002/nla

5. A LOWER BOUND ON $\mu(y,\gamma,\theta)$

The formula (13) for $\mu(y,\gamma,\theta)$ involves the smallest singular value of the $m\times(m+n+1)$ matrix N, which is expensive to compute directly. In this section we suggest a (hopefully) good lower bound on $\mu(y,\gamma,\theta)$, which can easily be estimated.

Theorem 5.1
With the notation and conditions of Theorem 3.1,

$$\mu(y,\gamma,\theta)\geqslant \min_{[\Delta A,\Delta b]\in\mathscr{C}^{+}_{A,b}(\gamma)} \|[\Delta A,\Delta b\,\theta]\|_2 \geqslant \mu^{lb}(y,\gamma,\theta)\equiv \frac{2\beta_0}{\sqrt{\beta_1^2+4\beta_0}+\beta_1} \tag{25}$$

where

$$\beta_0 \equiv \frac{\|(\gamma^{-2}+\|y\|^2)A^{\mathrm{T}}r+\|r\|^2 y\|}{(\gamma^{-2}+\|y\|^2)\sqrt{\theta^{-2}+\|y\|^2}+\|y\|(\theta^{-2}+\|y\|^2)} \tag{26}$$

$$\beta_1 \equiv \frac{(\gamma^{-2}+\|y\|^2)\sqrt{\theta^{-2}+\|y\|^2}\|A\|_2+(\gamma^{-2}+\|y\|^2)\|r\|+2\sqrt{\theta^{-2}+\|y\|^2}\|y\|\|r\|}{(\gamma^{-2}+\|y\|^2)\sqrt{\theta^{-2}+\|y\|^2}+\|y\|(\theta^{-2}+\|y\|^2)} \tag{27}$$

Proof
Obviously the first inequality in (25) holds. For any $[\Delta A,\Delta b]\in\mathscr{C}^{+}_{A,b}(\gamma)$, we have from (8) that

$$(\gamma^{-2}+\|y\|^2)\left(A+[\Delta A,\Delta b\,\theta]\begin{bmatrix}I\\0\end{bmatrix}\right)^{\mathrm{T}}\left(r+[\Delta A,\Delta b\,\theta]\begin{bmatrix}-y\\\theta^{-1}\end{bmatrix}\right)=-\left\|r+[\Delta A,\Delta b\,\theta]\begin{bmatrix}-y\\\theta^{-1}\end{bmatrix}\right\|^2 y$$

Denoting $F\equiv[\Delta A,\Delta b\,\theta]$ and $\alpha\equiv\gamma^{-2}+\|y\|^2$, we get

$$\begin{aligned}\alpha A^{\mathrm{T}}r+\|r\|^2 y=&-\alpha A^{\mathrm{T}}F\begin{bmatrix}-y\\\theta^{-1}\end{bmatrix}-\alpha[I,0]F^{\mathrm{T}}\left(r+F\begin{bmatrix}-y\\\theta^{-1}\end{bmatrix}\right)\\&-2r^{\mathrm{T}}F\begin{bmatrix}-y\\\theta^{-1}\end{bmatrix}y-[-y^{\mathrm{T}},\,\theta^{-1}]F^{\mathrm{T}}F\begin{bmatrix}-y\\\theta^{-1}\end{bmatrix}y\end{aligned}$$

Taking the 2-norm of both sides of this equation, we obtain the inequality

$$\begin{aligned}\|\alpha A^{\mathrm{T}}r+\|r\|^2 y\|\leqslant&\left(\alpha\|A\|_2\left\|\begin{bmatrix}-y\\\theta^{-1}\end{bmatrix}\right\|+\alpha\|r\|+2\|r\|\|y\|\left\|\begin{bmatrix}-y\\\theta^{-1}\end{bmatrix}\right\|\right)\|F\|_2\\&+\left(\alpha\left\|\begin{bmatrix}-y\\\theta^{-1}\end{bmatrix}\right\|+\|y\|\left\|\begin{bmatrix}-y\\\theta^{-1}\end{bmatrix}\right\|^2\right)\|F\|_2^2\end{aligned}$$

DOI: 10.1002/nla

that is, with (26) and (27), the quadratic inequality in terms of $\xi \equiv \|F\|_2$:

$$\beta_0 \leqslant \beta_1 \xi + \xi^2$$

Since ξ and β_0 are nonnegative, $\xi \geqslant \xi_+$, where ξ_+ is the positive root of $\beta_0 = \beta_1 \xi + \xi^2$, hence

$$\xi \geqslant \xi_+ = \frac{1}{2}\left(\sqrt{\beta_1^2 + 4\beta_0} - \beta_1\right) = 2\beta_0 \left(\sqrt{\beta_1^2 + 4\beta_0} + \beta_1\right)^{-1}$$

giving the second inequality in (25). □

The lower bound $\mu^{lb}(y, \gamma, \theta)$ in (25) can be estimated in $O(mn)$ flops, since $\|A\|_2$ can usually be estimated by a standard norm estimator in $O(mn)$ flops, see, for example, [14, Section 15.2]. In fact a good estimate of $\|A\|_2$ might already be available from whatever method is used for obtaining y, and in this case the cost will essentially be the $4mn$ flops required for computing $A^{\mathrm{T}}(b - Ay)$.

This bound can be compared with the lower bounds for the OLS minimal backward error such as those discussed in [9, 15]. If in (25) we take $\gamma \to 0$ to obtain a bound for the OLS problem and then take $\theta \to \infty$ to restrict the perturbations to A only, we obtain

$$\mu_{\mathrm{OLS}}(y, \infty) \equiv \lim_{\theta \to \infty} \mu_{\mathrm{OLS}}(y, \theta) = \lim_{\theta \to \infty} \lim_{\gamma \to 0} \mu(y, \gamma, \theta)$$

$$\geqslant \frac{2\|A^{\mathrm{T}} r\|}{\|A\|_2 \|y\| + \|r\| + \sqrt{(\|A\|_2 \|y\| + \|r\|)^2 + 4\|A^{\mathrm{T}} r\| \|y\|}}$$

This is exactly one of the bounds given in [9].

If in (25) we take $\gamma \to \infty$ to obtain a lower bound for the DLS extended minimal backward error and then take $\theta \to \infty$ to restrict the perturbations to A only, we obtain

$$\mu_{\mathrm{DLS}}(y, \infty) \equiv \lim_{\theta \to \infty} \mu_{\mathrm{DLS}}(y, \theta) = \lim_{\theta \to \infty} \lim_{\gamma \to \infty} \mu(y, \gamma, \theta) \geqslant \frac{2\hat{\beta}_0}{\sqrt{\hat{\beta}_1^2 + 4\hat{\beta}_0} + \hat{\beta}_1}$$

where

$$\hat{\beta}_0 = \frac{\left\| \|y\|^2 A^{\mathrm{T}} r + \|r\|^2 y \right\|}{2\|y\|^3}, \quad \hat{\beta}_1 = \frac{\|A\|_2 \|y\| + 3\|r\|}{2\|y\|}$$

This lower bound for the DLS problem was derived in [5].

6. AN ASYMPTOTIC ESTIMATE FOR $\mu(y, \gamma, \theta)$

Computing $\mu(y, \gamma, \theta)$ exactly is expensive and the lower bound (25) may not be very tight. In this section, by following [5, 16], we give an asymptotic estimate for $\mu(y, \gamma, \theta)$ and show how it is related to other asymptotic estimates in the literature for the OLS and DLS problems.

As in the previous sections, denote $r \equiv b - Ay$ and also $\alpha = \gamma^{-2} + \|y\|^2$. Let

$$h(A, b, y) \equiv A^{\mathrm{T}} r + \frac{\|r\|^2}{\gamma^{-2} + \|y\|^2} y = A^{\mathrm{T}} r + \|r\|^2 \alpha^{-1} y \tag{28}$$

DOI: 10.1002/nla

Recall from Section 3 that if $\hat{x}$ is the true STLS solution, $h(A,b,\hat{x})=0$. Then $\mu(y,\gamma,\theta)=\|[\widehat{\Delta A},\widehat{\Delta b}\,\theta]\|_F$, where $\{\widehat{\Delta A},\widehat{\Delta b}\}$ is the solution to $h(A+\Delta A,b+\Delta b,y)=0$ that minimizes $\|[\widehat{\Delta A},\widehat{\Delta b}\,\theta]\|_F$. By Taylor's expansion, for small enough $E\in\mathbb{R}^{m\times n}$ and $f\in\mathbb{R}^m$,

$$h(A+E,b+f,y)\approx h(A,b,y)+J_A\mathrm{vec}(E)+J_b f=h(A,b,y)+[J_A,\theta^{-1}J_b]\begin{bmatrix}\mathrm{vec}(E)\\ f\theta\end{bmatrix}$$

where $J_A\in\mathbb{R}^{n\times mn}$ and $J_b\in\mathbb{R}^{n\times m}$ are the Jacobian matrices of h with respect to $\mathrm{vec}(A)$ and b, respectively. Therefore, an approximation to $[\widehat{\Delta A},\widehat{\Delta b}\,\theta]$ is $[E,f\theta]$ such that

$$\|[E,f\theta]\|_F=\min\quad\text{s.t. } h(A,b,y)+[J_A,\theta^{-1}J_b]\begin{bmatrix}\mathrm{vec}(E)\\ f\theta\end{bmatrix}=0 \tag{29}$$

In other words,

$$\begin{bmatrix}\mathrm{vec}(E)\\ f\theta\end{bmatrix}=-[J_A,\theta^{-1}J_b]^{\dagger}h(A,b,y) \tag{30}$$

Thus, we obtain the following approximation to $\mu(y,\gamma,\theta)$:

$$\bar{\mu}(y,\gamma,\theta)\equiv\|[E,f\theta]\|_F=\left\|\begin{bmatrix}\mathrm{vec}(E)\\ f\theta\end{bmatrix}\right\|=\|[J_A,\theta^{-1}J_b]^{\dagger}h(A,b,y)\| \tag{31}$$

We say that $\bar{\mu}(y,\gamma,\theta)$ is an asymptotic estimate of $\mu(y,\gamma,\theta)$ due to the following result.

Theorem 6.1
Using the notation of Theorem 3.1 with $\mu(y,\gamma,\theta)$ defined as in (10) and $\bar{\mu}(y,\gamma,\theta)$ as in (31),

$$\lim_{y\to\hat{x}}\frac{\bar{\mu}(y,\gamma,\theta)}{\mu(y,\gamma,\theta)}=1$$

Proof
The proof is similar to that for Theorem 4.1 of [5]. A Taylor expansion of $h(A+\widehat{\Delta A},b+\widehat{\Delta b},y)$ gives

$$0=h(A+\widehat{\Delta A},b+\widehat{\Delta b},y)=h(A,b,y)+[J_A,\theta^{-1}J_b]\begin{bmatrix}\mathrm{vec}(\widehat{\Delta A})\\ \widehat{\Delta b}\,\theta\end{bmatrix}+\mathcal{O}(\|[\widehat{\Delta A},\widehat{\Delta b}]\|_F^2) \tag{32}$$

Substituting the resulting expression for $h(A,b,y)$ into (30) gives

$$\begin{bmatrix}\mathrm{vec}(E)\\ f\theta\end{bmatrix}=[J_A,\theta^{-1}J_b]^{\dagger}[J_A,\theta^{-1}J_b]\begin{bmatrix}\mathrm{vec}(\widehat{\Delta A})\\ \widehat{\Delta b}\,\theta\end{bmatrix}+\mathcal{O}(\|[\widehat{\Delta A},\widehat{\Delta b}]\|_F^2)$$

Taking the 2-norm of each side and noticing that $[J_A,\theta^{-1}J_b]^{\dagger}[J_A,\theta^{-1}J_b]$ is an orthogonal projection matrix, we obtain

$$\bar{\mu}(y,\gamma,\theta)\leqslant\mu(y,\gamma,\theta)+\mathcal{O}(\|[\widehat{\Delta A},\widehat{\Delta b}]\|_F^2) \tag{33}$$

DOI: 10.1002/nla

On the other hand, from (29) and (32) we obtain

$$[J_A, \theta^{-1}J_b]\begin{bmatrix}\mathrm{vec}(\widehat{\Delta A})\\ \widehat{\Delta b}\,\theta\end{bmatrix}=[J_A, \theta^{-1}J_b]\begin{bmatrix}\mathrm{vec}(E)\\ f\theta\end{bmatrix}+\mathcal{O}(\|[\widehat{\Delta A}, \widehat{\Delta b}]\|_F^2)$$

Since $[J_A, \theta^{-1}J_b]$ has full row rank (this will be demonstrated below), we can write

$$[J_A, \theta^{-1}J_b]\begin{bmatrix}\mathrm{vec}(\widehat{\Delta A})\\ \widehat{\Delta b}\,\theta\end{bmatrix}=[J_A, \theta^{-1}J_b]\left(\begin{bmatrix}\mathrm{vec}(E)\\ f\theta\end{bmatrix}+\mathcal{O}(\|[\widehat{\Delta A}, \widehat{\Delta b}]\|_F^2)\right)$$

Since

$$\begin{bmatrix}\mathrm{vec}(\widehat{\Delta A})\\ \widehat{\Delta b}\,\theta\end{bmatrix}$$

is the vector satisfying the above equality with minimum 2-norm, we must have

$$\mu(y,\gamma,\theta)\leqslant\bar{\mu}(y,\gamma,\theta)+\mathcal{O}(\|[\widehat{\Delta A}, \widehat{\Delta b}]\|_F^2) \tag{34}$$

The result follows from (33) and (34) with Corollary 3.1. □

In the following, we derive an explicit expression for $\bar{\mu}(y,\gamma,\theta)$. Write $A=[a_1,\ldots,a_n]$ and $y^{\mathrm{T}}=[\eta_1,\ldots,\eta_n]$. Given two column vectors $f\in\mathbb{R}^m$ and $g\in\mathbb{R}^n$, define the matrix $\partial f/\partial g^{\mathrm{T}}\equiv(\partial f_i/\partial g_j)\in\mathbb{R}^{m\times n}$. With this notation, using (28) we obtain

$$J_b=\frac{\partial h}{\partial b^{\mathrm{T}}}=A^{\mathrm{T}}+2\alpha^{-1}yr^{\mathrm{T}},\quad J_A=\begin{bmatrix}\dfrac{\partial h}{\partial a_1^{\mathrm{T}}},\ldots,\dfrac{\partial h}{\partial a_n^{\mathrm{T}}}\end{bmatrix} \tag{35}$$

Note that

$$\frac{\partial r}{\partial a_k^{\mathrm{T}}}=-\eta_k I,\quad \frac{\partial\|r\|^2}{\partial a_k^{\mathrm{T}}}=2r^{\mathrm{T}}\frac{\partial r}{\partial a_k^{\mathrm{T}}}=-2\eta_k r^{\mathrm{T}},\quad \frac{\partial(a_i^{\mathrm{T}}r)}{\partial a_k^{\mathrm{T}}}=\delta_{ik}r^{\mathrm{T}}-\eta_k a_i^{\mathrm{T}}$$

where $\delta_{ik}=1$ if $i=k$ and $\delta_{ik}=0$ otherwise. This gives

$$\frac{\partial h}{\partial a_k^{\mathrm{T}}}=e_k r^{\mathrm{T}}-\eta_k A^{\mathrm{T}}-2\alpha^{-1}\eta_k yr^{\mathrm{T}} \tag{36}$$

From (35) and (36), we obtain

$$\begin{aligned}
J_bJ_b^{\mathrm{T}}&=A^{\mathrm{T}}A+2\alpha^{-1}(A^{\mathrm{T}}ry^{\mathrm{T}}+yr^{\mathrm{T}}A)+4\|r\|^2\alpha^{-2}yy^{\mathrm{T}}\\
J_AJ_A^{\mathrm{T}}&=\sum_{k=1}^{n}(e_kr^{\mathrm{T}}-\eta_kA^{\mathrm{T}}-2\alpha^{-1}\eta_kyr^{\mathrm{T}})(e_kr^{\mathrm{T}}-\eta_kA^{\mathrm{T}}-2\alpha^{-1}\eta_kyr^{\mathrm{T}})^{\mathrm{T}}\\
&=\|y\|^2A^{\mathrm{T}}A+\alpha^{-1}(2\|y\|^2-\alpha)(A^{\mathrm{T}}ry^{\mathrm{T}}+yr^{\mathrm{T}}A)+\|r\|^2I+4\|r\|^2\alpha^{-2}(\|y\|^2-\alpha)yy^{\mathrm{T}}\\
&=\|y\|^2A^{\mathrm{T}}A+\alpha^{-1}(\|y\|^2-\gamma^{-2})(A^{\mathrm{T}}ry^{\mathrm{T}}+yr^{\mathrm{T}}A)+\|r\|^2I-4\|r\|^2\alpha^{-2}\gamma^{-2}yy^{\mathrm{T}}
\end{aligned}$$

DOI: 10.1002/nla

Denoting $J \equiv [J_A, \theta^{-1} J_b]$ and introducing the scalars

$$\xi_0 \equiv \sqrt{\|y\|^2 + \theta^{-2}}, \quad \xi_1 \equiv \frac{\|y\|^2 - \gamma^{-2} + 2\theta^{-2}}{\alpha \xi_0} \tag{37}$$

it follows that

$$\begin{aligned}
JJ^{\mathrm{T}} &\equiv J_A J_A^{\mathrm{T}} + \theta^{-2} J_b J_b^{\mathrm{T}} \\
&= \xi_0^2 A^{\mathrm{T}} A + \xi_0 \xi_1 (A^{\mathrm{T}} r y^{\mathrm{T}} + y r^{\mathrm{T}} A) + \|r\|^2 I + 4\|r\|^2 \alpha^{-2} (\theta^{-2} - \gamma^{-2}) y y^{\mathrm{T}} \\
&= (\xi_0 A + \xi_1 r y^{\mathrm{T}})^{\mathrm{T}} (\xi_0 A + \xi_1 r y^{\mathrm{T}}) + \|r\|^2 I + [4\|r\|^2 \alpha^{-2} (\theta^{-2} - \gamma^{-2}) - \xi_1^2 \|r\|^2] y y^{\mathrm{T}} \\
&= (\xi_0 A + \xi_1 r y^{\mathrm{T}})^{\mathrm{T}} (\xi_0 A + \xi_1 r y^{\mathrm{T}}) + \|r\|^2 I - \|r\|^2 \xi_0^{-2} y y^{\mathrm{T}} \\
&= (\xi_0 A + \xi_1 r y^{\mathrm{T}})^{\mathrm{T}} (\xi_0 A + \xi_1 r y^{\mathrm{T}}) + \|r\|^2 I + \|r\|^2 \|y\|^2 \xi_0^{-2} (I - y y^{\dagger}) - \|r\|^2 \|y\|^2 \xi_0^{-2} I \\
&= (\xi_0 A + \xi_1 r y^{\mathrm{T}})^{\mathrm{T}} (\xi_0 A + \xi_1 r y^{\mathrm{T}}) + \theta^{-2} \xi_0^{-2} \|r\|^2 I + \xi_0^{-2} \|r\|^2 \|y\|^2 (I - y y^{\dagger}) \\
&= \xi_0^2 B^{\mathrm{T}} B
\end{aligned}$$

where

$$B \equiv \begin{bmatrix} A + \xi_0^{-1} \xi_1 r y^{\mathrm{T}} \\ \xi_0^{-2} \|r\| \|y\| (I - y y^{\dagger}) \\ \theta^{-1} \xi_0^{-2} \|r\| I \end{bmatrix} = \begin{bmatrix} A + \dfrac{\gamma^2 \theta^2 \|y\|^2 - \theta^2 + 2\gamma^2}{(1 + \gamma^2 \|y\|^2)(1 + \theta^2 \|y\|^2)} r y^{\mathrm{T}} \\ \dfrac{\theta^2 \|r\| \|y\|}{1 + \theta^2 \|y\|^2} (I - y y^{\dagger}) \\ \dfrac{\theta \|r\|}{1 + \theta^2 \|y\|^2} I \end{bmatrix} \tag{38}$$

Notice from the last block of B that B has full column rank. Therefore, $J \equiv [J_A, \theta^{-1} J_b]$ has full row rank.

Defining

$$c \equiv \begin{bmatrix} \dfrac{\theta}{\sqrt{1 + \theta^2 \|y\|^2}} r \\ 0 \\ \dfrac{\|r\| (\theta^2 - \gamma^2)}{(1 + \gamma^2 \|y\|^2) \sqrt{1 + \theta^2 \|y\|^2}} y \end{bmatrix} \tag{39}$$

it is straightforward (but tedious) to verify that $B^{\mathrm{T}} c = \xi_0^{-1} h(A, b, y)$. Using (31), we obtain

$$\begin{aligned}
\bar{\mu}(y, \gamma, \theta) &= \|J^{\dagger} h(A, b, y)\| = \|J^{\mathrm{T}} (J J^{\mathrm{T}})^{-1} h(A, b, y)\| = \|(J J^{\mathrm{T}})^{-1/2} h(A, b, y)\| \\
&= \|(\xi_0^2 B^{\mathrm{T}} B)^{-1/2} h(A, b, y)\| = \|B (B^{\mathrm{T}} B)^{-1} B^{\mathrm{T}} c\| = \|B B^{\dagger} c\|
\end{aligned} \tag{40}$$

DOI: 10.1002/nla

Notice that $BB^\dagger$ is an orthogonal projector onto the range of B and $\bar{\mu}(y,\gamma,\theta)$ can be computed by using the QR factorization of B. This computation is more efficient than computing the SVD of N in (11) to obtain $\mu(y,\gamma,\theta)$.

The asymptotic estimate $\bar{\mu}(y,\gamma,\theta)$ is analogous to an estimate of the minimal backward error for the OLS problem whose various forms have been studied in [16–19], as well as for the DLS problem, derived in [5]. In the following sections, we show exactly how these are related.

6.1. Relationship to an OLS asymptotic estimate

Theorem 4.4 in [16] gives an asymptotic estimate for the minimal backward error for the OLS problem when only A is perturbed. This is extended in [17] and [19, Section 2.7] to the case when both perturbations to A and b are allowed. Using our notation, with ξ_0 defined in (37), the estimate is

$$\bar{\mu}_{\mathrm{OLS}}(y,\theta) \equiv \|(\xi_0^2 A^{\mathrm{T}}A + \|r\|^2 I)^{-1/2} A^{\mathrm{T}} r\| \tag{41}$$

It has the property that $\lim_{y\to\hat{x}} \bar{\mu}_{\mathrm{OLS}}(y,\theta)/\mu_{\mathrm{OLS}}(y,\theta) = 1$.

To compare our estimate $\bar{\mu}(y,\gamma,\theta)$ with that in (41) we need to take the limit as $\gamma \to 0$ to specialize our result to the OLS problem. It can be verified from (38) and (28) that

$$\lim_{\gamma\to 0} \xi_0^2 B^{\mathrm{T}}B = \xi_0^2 A^{\mathrm{T}}A - yr^{\mathrm{T}}A - A^{\mathrm{T}}ry^{\mathrm{T}} + \|r\|^2 I, \quad \lim_{\gamma\to 0} h(A,b,y) = A^{\mathrm{T}}r$$

Hence, from (40)

$$\lim_{\gamma\to 0} \bar{\mu}(y,\gamma,\theta) = \|(\xi_0^2 A^{\mathrm{T}}A - yr^{\mathrm{T}}A - A^{\mathrm{T}}ry^{\mathrm{T}} + \|r\|^2 I)^{-1/2} A^{\mathrm{T}} r\| \tag{42}$$

Notice that the terms involving $A^{\mathrm{T}}ry^{\mathrm{T}}$ in (42) are not present in (41). However, in the OLS problem, $A^{\mathrm{T}}r \to 0$ when $y \to \hat{x}$. Therefore, in the limit as $\gamma \to 0$ our estimate $\bar{\mu}(y,\gamma,\theta)$ in (40) is asymptotically equivalent to $\bar{\mu}_{\mathrm{OLS}}(y,\theta)$ in (41).

6.2. Relationship to a DLS asymptotic estimate

Section 4 in [5] introduced the following asymptotic estimate of the DLS extended minimal backward error with perturbations only in A

$$\bar{\mu}_{\mathrm{DLS}}(y,\infty) \equiv \|B_{\mathrm{DLS}} B_{\mathrm{DLS}}^{\dagger} c_{\mathrm{DLS}}\| \tag{43}$$

where

$$B_{\mathrm{DLS}} \equiv \begin{bmatrix} A + ry^\dagger \\ \|r\|/\|y\|(I - yy^\dagger) \end{bmatrix}, \quad c_{\mathrm{DLS}} \equiv \begin{bmatrix} r/\|y\| \\ 0 \end{bmatrix}$$

This estimate $\bar{\mu}_{\mathrm{DLS}}(y,\infty)$ has the property that $\lim_{y\to\hat{x}} \bar{\mu}_{\mathrm{DLS}}(y,\infty)/\mu_{\mathrm{DLS}}(y,\infty) = 1$.

To compare our estimate $\bar{\mu}(y,\gamma,\theta)$ with that in (43) we first take the limit as $\gamma \to \infty$ to specialize our result to the DLS problem, and then take the limit as $\theta \to \infty$ to restrict the perturbations to A

DOI: 10.1002/nla

only. From (38) and (39), when $\gamma \to \infty$ and then $\theta \to \infty$, we obtain

$$\lim_{\theta\to\infty}\lim_{\gamma\to\infty} B = \begin{bmatrix} A + ry^{\dagger} \\ \|r\|/\|y\|(I - yy^{\dagger}) \\ 0 \end{bmatrix} = \begin{bmatrix} B_{\mathrm{DLS}} \\ 0 \end{bmatrix}, \quad \lim_{\theta\to\infty}\lim_{\gamma\to\infty} c = \begin{bmatrix} r/\|y\| \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} c_{\mathrm{DLS}} \\ 0 \end{bmatrix}$$

It immediately follows from (40) that $\lim_{\theta\to\infty}\lim_{\gamma\to 0}\bar{\mu}(y,\gamma,\theta) = \bar{\mu}_{\mathrm{DLS}}(y,\infty)$.

7. NUMERICAL EXAMPLES

To illustrate our results we provide some numerical examples.

Recall from Section 3 that the extended minimal backward error $\mu(y,\gamma,\theta)$ in (10) is a lower bound on the true minimal backward error. Furthermore, from Theorem 3.2, if y is close enough to the true solution $\hat{x}$, then $\mu(y,\gamma,\theta)$ is in fact the true minimal backward error. We will test how close y must be to $\hat{x}$ for this to be the case. We can verify whether or not $\mu(y,\gamma,\theta)$ is the true minimal backward error as follows. If the inequality in (7) holds with the optimal $\widehat{\Delta A}$ and $\widehat{\Delta b}$ in (14) and (15), then minimizing over $\mathscr{C}^{+}_{A,b}(\gamma)$ in (8) gives the same solution as minimizing over $\mathscr{C}_{A,b}(\gamma)$ in (7), hence $\mu(y,\gamma,\theta)$ is indeed the true minimal backward error.

We performed our numerical examples in MATLAB version 7.4.0 on a 3.20 GHz Intel Pentium 4 processor running Gentoo Linux. The functions randn and rand below are MATLAB built-in functions that generate matrices whose elements are sampled from normal and uniform distributions, respectively. We create our test problems as follows:

- We create two types of matrices $A \in \mathbb{R}^{120\times 80}$:
  Type 1: $A = \tilde{A}/\|\tilde{A}\|_F$, $\tilde{A} = \mathrm{randn}(120, 80)$. Usually $\kappa_2(A) \leqslant 10$.
  Type 2: $A = \tilde{A}/\|\tilde{A}\|_F$, $\tilde{A} = U\Sigma V^{\mathrm{T}}$, where U and V are the Q-factors of the QR factorization of random matrices $\mathrm{rand}(120, 120)$ and $\mathrm{rand}(80, 80)$, respectively, and $\Sigma = \mathrm{diag}(\sigma_i)$ with logarithmically equally spaced singular values σ_i between 10^0 and 10^{-4}. Note that here $\kappa_2(A) = 10^4$.
- In both cases we create b as follows: $b = Ax$, where $x = [1, \ldots, 1]^{\mathrm{T}}$.
- We let $A = A + (1/(\sqrt{120\times 80}))\delta_A \cdot \mathrm{rand}(120, 80)$ and $b = b + (1/(\sqrt{120}))\delta_b\|b\| \cdot \mathrm{rand}(120, 1)$ so that the system is likely no longer compatible, with $\delta_A = \delta_b = 10^{-6}, \ldots, 10^{-1}$.
- For simplicity we use $\gamma = \theta = 1$. Results with other parameters γ and θ are very similar.
- We compute the STLS solution $\hat{x}$ by using the SVD of $[A, b\gamma]$ (see [6]) and let the approximate solution y be

$$y = \hat{x} + \frac{1}{\sqrt{80}}\delta_{\hat{x}}\|\hat{x}\| \cdot \mathrm{rand}(80, 1)$$

  with $\delta_x = 0, 10^{-6}, \ldots, 10^{-1}$.
- For each pair of δ_A and $\delta_{\hat{x}}$ and each type of matrix, we generate 1000 sample problems.

In our numerical tests, we use single precision to generate the data and to compute the STLS solution $\hat{x}$. When verifying if the inequality in (7) holds with the optimal $\widehat{\Delta A}$ and $\widehat{\Delta b}$ in (14) and (15), we must take into account the fact that the computations involved in verifying the inequality are themselves subject to rounding errors. To see what effect this has on our results, we first use

DOI: 10.1002/nla

single precision to generate the data A and b, and to compute $\hat{x}$ and y. We then use both single and then double precision to compute the quantities

$$\frac{\|b+\widehat{\Delta b}-(A+\widehat{\Delta A})y\|^2}{\gamma^{-2}+\|y\|^2} \quad \text{and} \quad \sigma_{\min}^2(A+\widehat{\Delta A})$$

The number of failures to satisfy the inequality in (7) with optimal $\widehat{\Delta A}$ and $\widehat{\Delta b}$, when verified using single precision (S) and double precision (D), are presented in Tables I and II.

Tables I and II indicate that the inequality almost always holds when it is verified using double precision, and holds less frequently for Type 2 matrices that are more ill-conditioned when it is verified using single precision. This strongly suggests that the times in which the inequality failed to hold in Tables I and II were likely due to rounding errors when actually verifying the inequality.

Table I. Number of failures to satisfy the inequality in (7) out of 1000 samples for Type 1 A.

			$\delta_{\hat{x}}$						
			0	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
$\delta_A,\ \delta_b$	10^{-6}	S	0	0	0	0	0	0	0
		D	0	0	0	0	0	0	0
	10^{-5}	S	0	0	0	0	0	0	0
		D	0	0	0	0	0	0	0
	10^{-4}	S	0	0	0	0	0	0	0
		D	0	0	0	0	0	0	0
	10^{-3}	S	0	0	0	0	0	0	0
		D	0	0	0	0	0	0	0
	10^{-2}	S	0	0	0	0	0	0	0
		D	0	0	0	0	0	0	0
	10^{-1}	S	0	0	1	0	0	1	3
		D	0	0	0	0	0	0	3

Table II. Number of failures to satisfy the inequality in (7) out of 1000 samples for Type 2 A.

			$\delta_{\hat{x}}$						
			0	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
$\delta_A,\ \delta_b$	10^{-6}	S	0	0	0	0	0	0	0
		D	0	0	0	0	0	0	0
	10^{-5}	S	0	0	0	0	0	0	0
		D	0	0	0	0	0	0	0
	10^{-4}	S	0	0	0	0	0	0	0
		D	0	0	0	0	0	0	0
	10^{-3}	S	16	12	9	6	11	16	20
		D	0	0	0	0	0	0	13
	10^{-2}	S	7	11	8	5	4	8	14
		D	0	0	0	0	0	0	7
	10^{-1}	S	1	1	2	0	4	4	12
		D	0	0	0	0	0	0	10

DOI: 10.1002/nla

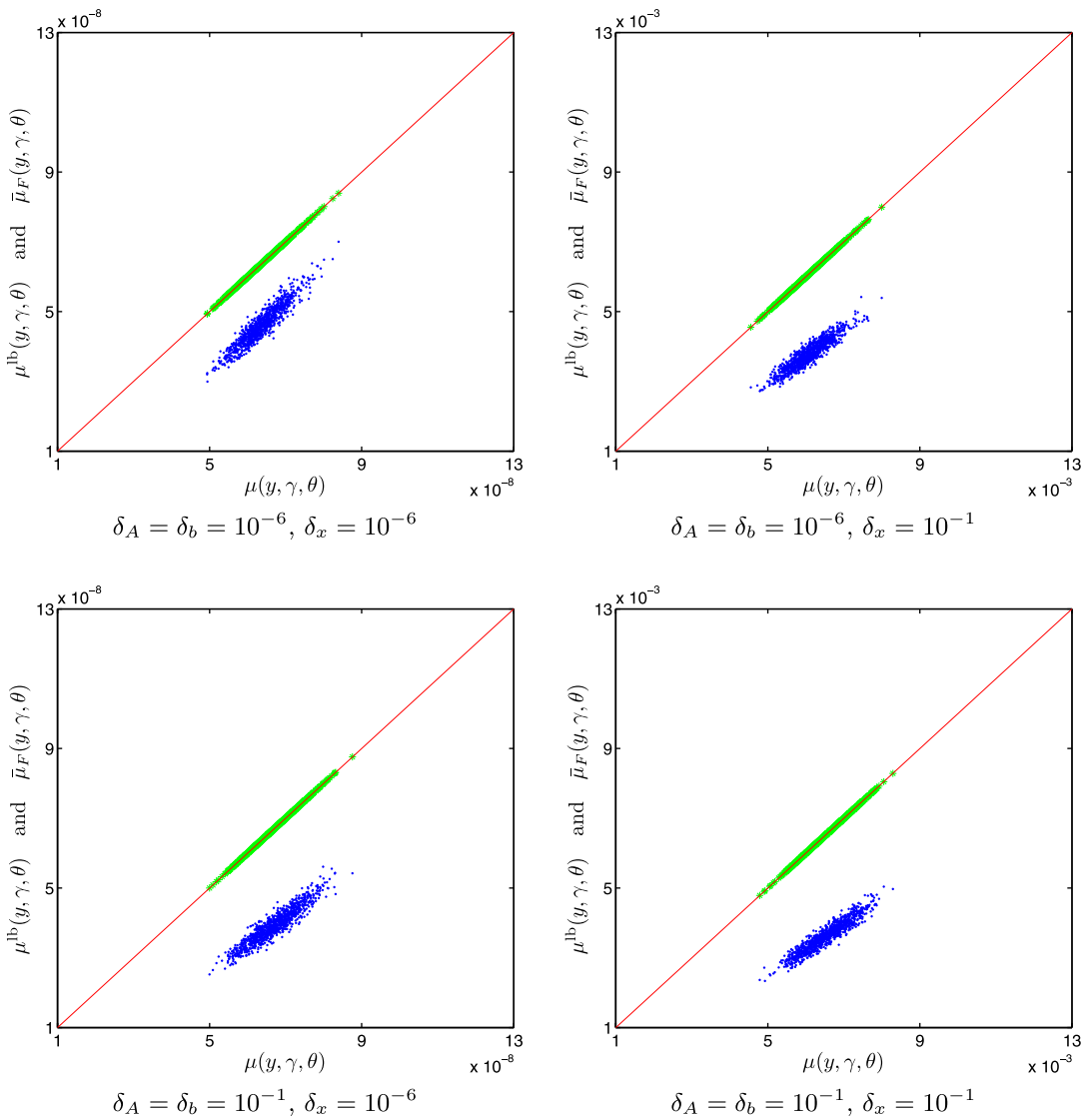

Figure 1. Type 1 A: $\mu^{lb}(y,\gamma,\theta)$ (dots) and $\bar{\mu}(y,\gamma,\theta)$ (stars) versus $\mu(y,\gamma,\theta)$.

Note that $\hat{x}$ in our tests is not the exact STLS solution but a computed solution (computed by a numerically reliable algorithm). Therefore, each y is not a perturbation of the true STLS solution but rather of a computed solution. However, we note that repeating the above experiment with $\hat{x}$ computed in double precision gave almost identical results. We conclude that when the approximate solution y is a reasonable approximation to the true STLS solution, it is usually reasonable to expect that $\mu(y,\gamma,\theta)$ in (10) is indeed the true minimal backward error.

The formula for $\mu(y,\gamma,\theta)$ in (13) involves the smallest singular value of an $m\times(m+n+1)$ matrix, which is expensive to compute directly. In Sections 5 and 6 we gave two estimates for

DOI: 10.1002/nla

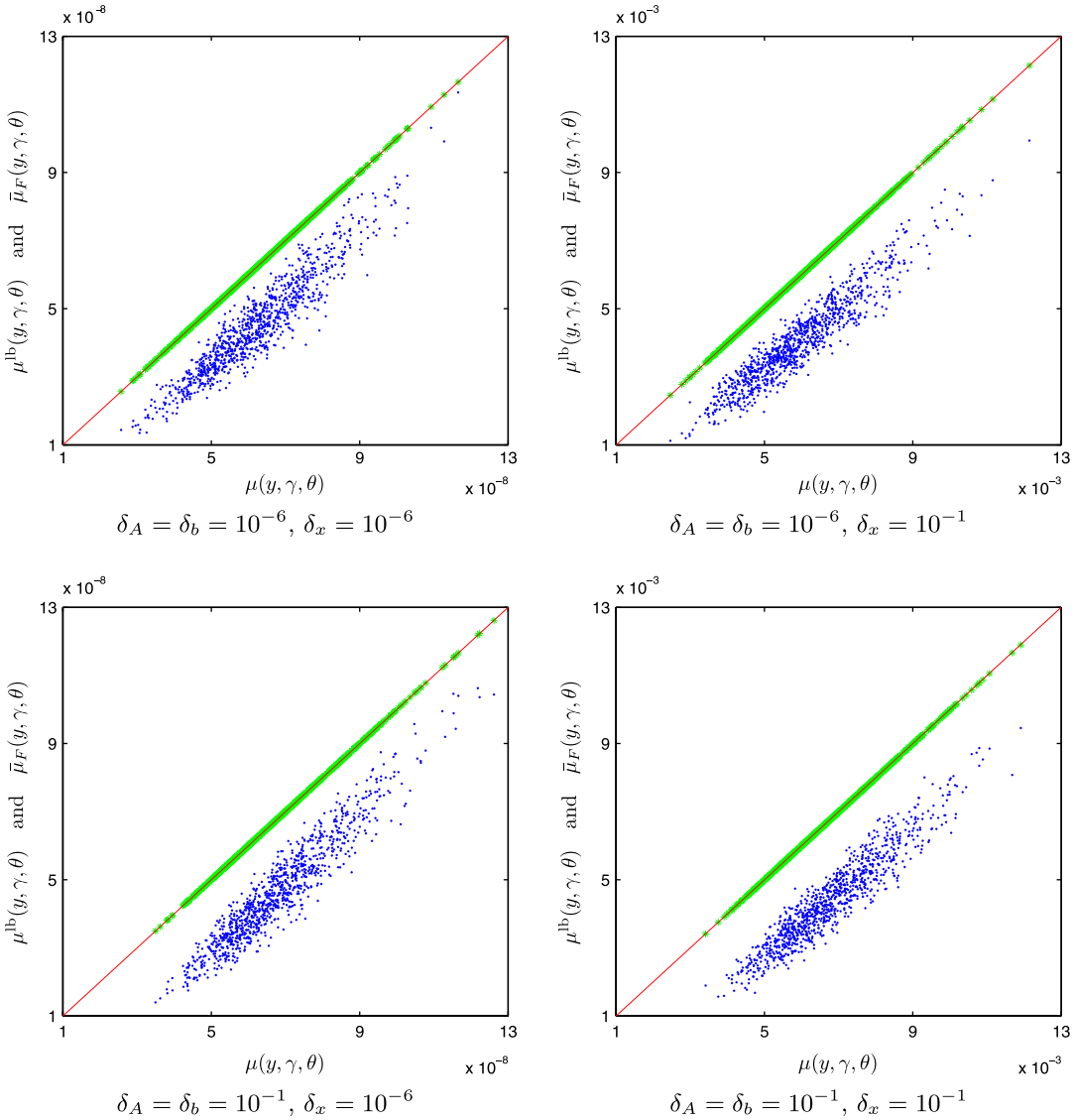

Figure 2. Type 2 A: $\mu^{lb}(y,\gamma,\theta)$ (dots) and $\bar{\mu}(y,\gamma,\theta)$ (stars) versus $\mu(y,\gamma,\theta)$.

$\mu(y,\gamma,\theta)$: the lower bound $\mu^{lb}(y,\gamma,\theta)$ in (25) and the asymptotic estimate $\bar{\mu}(y,\gamma,\theta)$ in (40). We will test how good an approximation of these quantities are to $\mu(y,\gamma,\theta)$. In Figures 1 and 2 we plot the lower bound $\mu^{lb}(y,\gamma,\theta)$ versus $\mu(y,\gamma,\theta)$ in dots and the asymptotic estimate $\bar{\mu}(y,\gamma,\theta)$ versus $\mu(y,\gamma,\theta)$ in stars. The diagonal line is plotted for reference. Here, we show four cases for each test problem, with all quantities computed in double precision. Results for all the other test problems were very similar.

DOI: 10.1002/nla

These (as well as other) examples suggest that the lower bound $\mu^{lb}(y,\gamma,\theta)$ gives good order of magnitude estimates of $\mu(y,\gamma,\theta)$, while the asymptotic estimate $\bar{\mu}(y,\gamma,\theta)$ gives an excellent estimate of $\mu(y,\gamma,\theta)$.

8. CONCLUSIONS

Given an approximate STLS solution y, we have found an expression for an extended minimal backward error $\mu(y,\gamma,\theta)$ in (10). This is an asymptotically tight lower bound on the true minimal backward error. Our numerical tests suggest that when the approximate STLS solution y is a reasonable approximation to the true STLS solution $\hat{x}$, $\mu(y,\gamma,\theta)$ is the true minimal backward error. We therefore believe that $\mu(y,\gamma,\theta)$ can be used in practice as a measure of backward error.

Since $\mu(y,\gamma,\theta)$ is very expensive to compute directly, we have given a lower bound on it, $\mu^{lb}(y,\gamma,\theta)$ in (25), as well as an asymptotic estimate for it, $\bar{\mu}(y,\gamma,\theta)$ in (40). In all our numerical tests the lower bound $\mu^{lb}(y,\gamma,\theta)$ gives good order of magnitude estimates of $\mu(y,\gamma,\theta)$, while the asymptotic estimate $\bar{\mu}(y,\gamma,\theta)$ gives an excellent estimate of $\mu(y,\gamma,\theta)$. The lower bound can be computed very efficiently. In the future we intend to find efficient and reliable ways to estimate the asymptotic estimate $\bar{\mu}(y,\gamma,\theta)$.

The main contribution of this paper is to provide a backward perturbation analysis for the STLS problem, and in so doing unify the backward perturbation analyses for OLS, DLS and TLS problems. In the extreme cases as $\gamma \to 0$ and $\gamma \to \infty$, the STLS problem becomes the OLS and DLS problems, respectively. We have shown how $\mu(y,\gamma,\theta)$, its lower bound $\mu^{lb}(y,\gamma,\theta)$ and its asymptotic estimate $\bar{\mu}(y,\gamma,\theta)$ specialize to corresponding estimates of backward error in the literature for the OLS and DLS problems.

ACKNOWLEDGEMENTS

We are grateful to Chris Paige for his many valuable suggestions. We also thank two referees for their helpful comments.

REFERENCES

1. Arioli M, Duff I, Ruiz D. Stopping criteria for iterative solvers. *SIAM Journal on Matrix Analysis and Applications* 1992; **13**:138–144.
2. Chang X-W, Paige CC, Titley-Peloquin D. Stopping criteria for the iterative solution of linear least squares problems. *SIAM Journal on Matrix Analysis and Applications*, to appear.
3. Paige CC, Saunders MA. LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software* 1982; **8**:43–71.
4. Rigal JL, Gaches J. On the compatibility of a given solution with the data of a linear system. *Journal of the ACM* 1967; **14**:543–548.
5. Chang X-W, Golub GH, Paige CC. Towards a backward perturbation analysis for data least squares problems. *SIAM Journal on Matrix Analysis and Applications* 2008; **30**(4):1281–1301.
6. Paige CC, Strakoš Z. Scaled total least squares fundamentals. *Numerische Mathematik* 2002; **91**:117–146.
7. Paige CC, Strakoš Z. Unifying least squares, total least squares and data least squares. In *Total Least Squares and Errors-in-Variables Modeling*, Van Huffel S, Lemmerling P (eds). Kluwer Academic Publishers: Dordrecht, 2002; 25–34.
8. Rao BD. Unified treatment of LS, TLS and truncated SVD methods using a weighted TLS framework. In *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modelling*, Van Huffel S (ed.). SIAM: Philadelphia, PA, 1997; 11–20.

DOI: 10.1002/nla

9. Waldén B, Karlson R, Sun J-G. Optimal backward perturbation bounds for the linear least squares problem. *Numerical Linear Algebra with Applications* 1995; **2**:271–286.
10. Golub GH, Van Loan CF. An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis* 1980; **17**:883–893.
11. van Huffel S, Vandewalle J. *The Total Least Squares Problem*: *Computational Aspects and Analysis*. SIAM: Philadelphia, PA, 1991.
12. Chang X-W, Paige CC, Titley-Peloquin D. Characterizing matrices that are consistent with given solutions. *SIAM Journal on Matrix Analysis and Applications* 2008; **30**(4):1406–1420.
13. Horn RA, Johnson CR. *Matrix Analysis*. Cambridge University Press: New York, NY, 1985.
14. Higham NJ. *Accuracy and Stability of Numerical Algorithms* (2nd edn). SIAM: Philadelphia, PA, 2002.
15. Karlson R, Waldén B. Estimation of backward perturbation bounds for the linear least squares problem. *BIT* 1997; **37**:862–869.
16. Grcar JF. Optimal sensitivity analysis of linear least squares. *Technical Report LBNL-52434*, Lawrence Berkeley National Laboratory, 2003.
17. Grcar JF, Saunders MA, Su Z. Estimates of optimal backward perturbations for linear least squares problems. Manuscript, 2004.
18. Gu M. Backward perturbation bounds for linear least squares problems. *SIAM Journal on Matrix Analysis and Applications* 1998; **20**:363–372.
19. Su Z. Computational methods for least squares problems and clinical trials. *Ph.D. Thesis*, Scientific Computing and Computational Mathematics, Stanford University, 2005.

DOI: 10.1002/nla