TOWARDS A BACKWARD PERTURBATION ANALYSIS FOR DATA LEAST SQUARES PROBLEMS

X.-W. CHANG*, G. H. GOLUB†, AND C.C. PAIGE‡

Xiao-Wen Chang and Chris Paige dedicate this in memory of their warm, generous and inspirational friend Gene Golub.

Abstract. Given an approximate solution to a data least squares (DLS) problem, we would like to know its minimal backward error. Here we derive formulas for what we call an "extended" minimal backward error, which is at worst a lower bound on the minimal backward error. When the given approximate solution is a good enough approximation to the exact solution of the DLS problem (which is the aim in practice), the extended minimal backward error is the actual minimal backward error, and this is also true in other easily assessed and common cases. Since it is computationally expensive to compute the extended minimal backward error directly, we derive a lower bound on it and an asymptotic estimate for it, both of which can be evaluated less expensively. Simulation results show that for reasonable approximate solutions the lower bound has the same order as the extended minimal backward error, and the asymptotic estimate is an excellent approximation to the extended minimal backward error.

Key words. data least squares, backward errors, numerical stability, perturbation analysis, asymptotic estimate, iterative methods, stopping criteria.

AMS subject classifications. 15A06, 65F20, 65G50.

1. Introduction. Given an approximate solution to a problem, the aim of backward perturbation analysis is to find a minimum size perturbation in the data such that the approximate solution is an exact solution of the perturbed problem. In the analysis one tries to find a formula for, or good bounds on, the size of the minimal perturbation (to be referred to as the minimal backward error) and design an efficient algorithm to evaluate or estimate the formula or the bounds. If the relative minimal backward error (i.e., the size of the minimal perturbation divided by an acceptable measure of the size of the data) is of the order of the unit round-off then we say that the approximate solution is a (normwise) backward stable solution. Backward perturbation analyses are useful in practice. Sometimes we may not know if an algorithm for solving a problem is numerically stable, e.g., the backward numerical stability of some fast algorithms for structured matrix problems is unknown. But if we know that a computed solution of a specific problem is a backward stable solution, we are satisfied with this computed solution. Also when we solve a large scale problem by an iterative algorithm, the results of a backward perturbation analysis can often be used to design effective stopping criteria, see, for example, [1], [20] and [25].

There has been a lot of work on the backward perturbation analysis of linear systems, especially in recent years. For example, for consistent linear systems, see [14], [25], [30], [31], [32], [34], [37]; for unconstrained least squares problems, see [9], [12], [17]–[19], [26]–[30], [35]; and for constrained least squares problems, see [4], [18] and [19].

*School of Computer Science, McGill University, Montreal, Quebec, Canada, H3A 2A7. Email: chang@cs.mcgill.ca. The work of this author was supported by NSERC of Canada Grant RGPIN217191-03.

†Born 29 February 1932, died 16 November 2007. Previously: Department of Computer Science, Stanford University, CA 94305-9025. His work was in part supported by an NSF grant QAACT.

‡School of Computer Science, McGill University, Montreal, Quebec, Canada, H3A 2A7. Email: paige@cs.mcgill.ca. The work of this author was supported by NSERC of Canada Grant RGPIN9236.

The main purpose of this paper is to give a norm-wise backward perturbation analysis for the general linear data least squares (DLS) problem. As a result the structure of the matrix and magnitudes of individual elements of the matrix in the DLS problem will not be considered. We derive formulas for an "extended" minimal backward error in section 2. This extended minimal backward error is at worst a lower bound on the minimal backward error. But we show that when the given approximate solution is a good enough approximation to the exact solution of the DLS problem (which is the aim in practice), the extended minimal backward error is the actual minimal backward error. Section 2.1 deals with perturbations in both A and b, while section 2.2 considers perturbations in A alone, and shows how these are limiting cases of those in section 2.1. Since computing the extended minimal backward error directly is time consuming, in section 3 we derive a lower bound on, and in section 4 an asymptotic estimate for, this extended minimal backward error. We give numerical examples in section 5. Finally a summary is given in section 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 column range is denoted by $\mathcal{R}(B)$, its Moore-Penrose generalized inverse is denoted by $B^\dagger$, its smallest singular value (the p-th largest singular value with $p=\min\{m,n\}$) by $\sigma_{\min}(B)$, and its condition number in the 2-norm is denoted by $\kappa_2(A)$. For any symmetric $B\in\mathbb{R}^{n\times n}$, its eigenvalues are labeled in non-decreasing order: $\lambda_{\min}=\lambda_1\le\lambda_2\le\cdots\le\lambda_n$, but when only $\lambda_{\min}$ is of interest we will write $\lambda=\lambda_{\min}$. For any vector $v\in\mathbb{R}^n$, its Moore-Penrose generalized inverse is

$$v^\dagger\equiv\begin{cases}0 & \text{if } v=0,\\ v^T/\|v\|_2^2 & \text{if } v\neq 0;\end{cases}\qquad \|v\|_2\equiv(v^Tv)^{\frac{1}{2}}.$$

Note that $vv^\dagger$ is the orthogonal projector onto $\mathcal{R}(v)$ and $I-vv^\dagger$ is the orthogonal projector onto the orthogonal complement of $\mathcal{R}(v)$.

2. Backward perturbation analysis. Given $A\in\mathbb{R}^{m\times n}$ and $b\in\mathbb{R}^m$, the data least squares (DLS) problem defined by DeGroat and Dowling [5] is:

$$\sigma_{D}\equiv\min_{E,x}\|E\|_F\quad\text{subject to}\quad(A+E)x=b,\qquad\|E\|_F\equiv[\mathrm{trace}(E^TE)]^{\frac{1}{2}}. \tag{2.1}$$

See also, for example, [21, 22]. The purpose of the DLS problem is to find the optimal x. For applications of the DLS method to some signal processing problems, see [5]. Let $\mathcal{U}_{min}(A)$ be the left singular vector subspace of A corresponding to its minimum singular value $\sigma_{min}(A)$. In [21] it was explained that a satisfactory condition for building the theory for the DLS problem (2.1) is the condition that we will now assume holds:

$$A\text{ has full column rank, and }b\not\perp\mathcal{U}_{min}(A). \tag{2.2}$$

With this condition the solution to (2.1) must exist and be unique. From $(A+E)x=b$ we have $Ex=b-Ax$. Thus the minimal E must satisfy

$$E=(b-Ax)x^\dagger. \tag{2.3}$$

But (2.2) implies $b\neq 0$, so the solution x must be nonzero and this allows us to eliminate E and reformulate the DLS problem (2.1) as

$$\sigma_{D}\equiv\min_x\|(b-Ax)x^\dagger\|_F=\min_x\frac{\|b-Ax\|_2}{\|x\|_2}. \tag{2.4}$$

From [21, (5.14)–(5.17)], $\hat{x}$ solves the DLS problem (2.1) if and only if

$$A^T(b - A\hat{x}) = -\hat{x}\frac{\|b - A\hat{x}\|_2^2}{\|\hat{x}\|_2^2}, \tag{2.5}$$

$$\frac{\|b - A\hat{x}\|_2}{\|\hat{x}\|_2} < \sigma_{\min}(A). \tag{2.6}$$

Differentiating the objective function in (2.4) and setting the result to zero gives (2.5), corresponding to a stationary point. The global minimum also satisfies (2.6).

The DLS formulation is designed for problems where the right hand side b is accurately known, but the matrix A is only known approximately. Given a nonzero approximate solution $y \in \mathbb{R}^n$ to $Ax \approx b$, two questions are of particular interest here:

Q1: Is y a feasible (not necessarily DLS) solution, given the accuracy of the data?

Q2: Is y a backward stable solution to the DLS problem for the given data A, b?

Q1 will often be easy to check: for example if it is known that the given data matrix A approximates an unknown ideal matrix $\hat{A}$ to within $\|A - \hat{A}\|_{\{F \text{ or } 2\}} \le \alpha$ while b is accurately known, then from (2.3) we need only check that $\|b - Ay\|/\|y\| \le \alpha$. If the answer to Q1 is positive and we are not interested in the DLS solution, we might accept y. But then in practice there will be an infinite set of y satisfying Q1, and we will often seek some additional criterion, for example "does y make sense physically?"—a difficult question we might ask of an ill-posed problem. Here we consider the more generally approachable question Q2, since if we can answer this affirmatively we will know that y is a desirable *computational* solution to (2.1). Even if the answer to Q1 is "no" we might still check Q2, since it is possible for y to satisfy Q2 but not Q1, in that y can be a DLS solution for $A + \Delta A$, $b + \Delta b$ for very small ΔA and Δb, but the minimal norm E in $(A + \Delta A + E)y = b + \Delta b$ can be too large for Q1. This would indicate that there are difficulties with the data.

To answer Q2 we would like to solve the minimal backward error problem:

$$\min_{\Delta A, \Delta b} \|[\Delta A, \Delta b\, \theta]\|_F \quad \text{subject to} \quad y = \arg\min_x \frac{\|b + \Delta b - (A + \Delta A)x\|_2}{\|x\|_2}, \tag{2.7}$$

see (2.4), where the chosen scalar $\theta \ge 0$ allows a different emphasis on each data error.

From (2.5) and (2.6) we see that $[\Delta A, \Delta b]$ is a backward perturbation for the DLS problem with the given solution y if and only if it is in the set $\mathcal{C}_{A,b}$ where

$$\mathcal{C}^+_{A,b} \equiv \big\{[\Delta A, \Delta b] :\ (A+\Delta A)^T[b+\Delta b - (A+\Delta A)y] = -y\frac{\|b+\Delta b - (A+\Delta A)y\|_2^2}{\|y\|_2^2}\big\}, \tag{2.8}$$

$$\mathcal{C}_{A,b} \equiv \big\{[\Delta A, \Delta b] :\ [\Delta A, \Delta b] \in \mathcal{C}^+_{A,b}\ \&\ \frac{\|b+\Delta b - (A+\Delta A)y\|_2}{\|y\|_2} < \sigma_{min}(A+\Delta A)\big\}. \tag{2.9}$$

The inequality in (2.9) makes it difficult to derive a general expression for $[\Delta A, \Delta b] \in \mathcal{C}_{A,b}$, so we initially ignore it and consider the larger set $\mathcal{C}^+_{A,b}$, which we will show is also useful. The following result from Theorem 5.1 of [3] characterizes $[\Delta A, \Delta b] \in \mathcal{C}^+_{A,b}$.

LEMMA 2.1. *If $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $y \in \mathbb{R}^n$ is nonzero, then $[\Delta A, \Delta b] \in \mathcal{C}^+_{A,b}$ in (2.8) if and only if there exist $w \in \mathbb{R}^n$ and $Z \in \mathbb{R}^{m\times n}$ such that*

$$A + \Delta A = (b + \Delta b - w)y^\dagger + (I - ww^\dagger)Z(I - yy^\dagger), \quad (b + \Delta b)^T w = 0. \tag{2.10}$$

2.1. Allowing backward perturbations in A and b. Based on Lemma 2.1 we will first find a computable expression for $\mu_F(y,\theta)$ in the following "extended" minimal backward error problem:

$$\mu_F(y,\theta) \equiv \min_{[\Delta A,\Delta b]\in\mathcal{C}^+_{A,b}} \|[\Delta A, \Delta b\,\theta]\|_F, \quad \text{for } \mathcal{C}^+_{A,b} \text{ in (2.8).} \tag{2.11}$$

We call $\mu_F(y,\theta)$ the extended minimal backward error because we minimize over the extended set $\mathcal{C}^+_{A,b}$, giving at worst a lower bound on the minimal backward error.

If $b=0$ the DLS problem (2.1) has the solution $\hat{x}=0$; if $y=0$ then it cannot be the DLS solution of any problem with $b\neq 0$. We do not need to consider these cases further. For the remainder of this paper we will assume the conditions and notation of the following theorem.

The following will simplify the presentation:

$$\rho \equiv 1/(1+\theta^2\|y\|_2^2), \qquad \text{so that} \quad \rho\theta^2\|y\|_2^2 = 1-\rho \quad \text{and} \quad 0\le\rho\le 1. \tag{2.12}$$

THEOREM 2.2. *Suppose that we are given $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$, nonzero $y\in\mathbb{R}^n$ and $\theta\ge 0$; and suppose that (2.2) holds. Let $r\equiv b-Ay$, $\rho\equiv 1/(1+\theta^2\|y\|_2^2)$, and*

$$N \equiv N(\theta) \equiv [A(I-yy^\dagger), \rho^{\frac{1}{2}}\theta\|r\|_2(I-rr^\dagger), b\,\theta], \tag{2.13}$$

$$M \equiv M(\theta) \equiv A\left(I-yy^\dagger\right)A^T - r\rho\theta^2 r^T + b\,\theta^2 b^T = NN^T - \rho\theta^2\|r\|_2^2 I. \tag{2.14}$$

Then $M(\theta)$ has at most one negative eigenvalue, and the DLS extended minimal backward error $\mu_F(y,\theta)$ in (2.11) satisfies

$$\mu_F^2(y,\theta) = \begin{cases} \rho\theta^2\|r\|_2^2 & \text{if } \lambda_{\min}(M(\theta))\ge 0, \\ \rho\theta^2\|r\|_2^2 + \lambda_{\min}(M(\theta)) = \sigma^2_{\min}(N(\theta)) & \text{if } \lambda_{\min}(M(\theta))<0. \end{cases} \tag{2.15}$$

Furthermore $\mu_F(y,\theta)$ is given by the backward perturbations $\widehat{\Delta A}$ and $\widehat{\Delta b}$ in

$$A+\widehat{\Delta A} = \begin{cases} A+r(1-\rho)y^\dagger & \text{if } \lambda_{\min}(M(\theta))\ge 0, \\ (I-w_\theta w_\theta^\dagger)[A+r(1-\rho)y^\dagger]+w_\theta w_\theta^\dagger Ayy^\dagger & \text{if } \lambda_{\min}(M(\theta))<0, \end{cases} \tag{2.16}$$

$$b+\widehat{\Delta b} = \begin{cases} b-r\rho & \text{if } \lambda_{\min}(M(\theta))\ge 0, \\ (I-w_\theta w_\theta^\dagger)(b-r\rho) & \text{if } \lambda_{\min}(M(\theta))<0, \end{cases} \tag{2.17}$$

where w_θ is the ('unique' when $\lambda_{\min}(M(\theta))<0$) eigenvector of $M(\theta)$ corresponding to $\lambda_{\min}(M(\theta))$, also the left singular vector of $N(\theta)$ corresponding to $\sigma_{\min}(N(\theta))$.

Proof. $A(I-yy^\dagger)A^T+b\,\theta^2 b^T$ is nonnegative definite, so from [15, Theorem 4.3.4(b)] with $k=1$, $M(\theta)$ has at most one negative eigenvalue.

Now we want to determine the optimal w, Z, ΔA and Δb in (2.10) to minimize $\|\Delta A, \Delta b\,\theta\|_F$. In the following we discuss two cases separately.

Case 1: The optimal $w=0$. Let $Y=[y/\|y\|_2, Y_2]\in\mathbb{R}^{n\times n}$ be an orthogonal matrix. From (2.10) we have $(b+\Delta b)^T w=0$ automatically, and

$$\begin{aligned} \Delta A Y &= (b+\Delta b)y^\dagger[y/\|y\|_2, Y_2] + Z(I-yy^\dagger)[y/\|y\|_2, Y_2] - A[y/\|y\|_2, Y_2] \\ &= [(b+\Delta b)/\|y\|_2, 0] + [0, ZY_2] - [Ay/\|y\|_2, AY_2] \\ &= [(r+\Delta b)/\|y\|_2, (Z-A)Y_2]. \end{aligned}$$

It follows that

$$\|[\Delta A, \Delta b\,\theta]\|_F^2 = \|\Delta A Y\|_F^2 + \theta^2\|\Delta b\|_2^2 = \frac{1}{\|y\|_2^2}\left\|\begin{bmatrix} I \\ \theta\|y\| I \end{bmatrix}\Delta b + \begin{bmatrix} r \\ 0 \end{bmatrix}\right\|_2^2 + \|(Z-A)Y_2\|_F^2.$$

Thus $\|[\Delta A, \Delta b\,\theta]\|_F$ is minimized when

$$\Delta b = \widehat{\Delta b} \equiv -\begin{bmatrix} I \\ \theta\|y\|I \end{bmatrix}^{\dagger} \begin{bmatrix} r \\ 0 \end{bmatrix} = -r\rho, \quad Z = \widehat{Z} \equiv A, \tag{2.18}$$

and from (2.10) we see that the optimal ΔA must satisfy

$$\Delta A = \widehat{\Delta A} \equiv (b - r\rho)y^{\dagger} + A(I - yy^{\dagger}) - A = r(1-\rho)y^{\dagger}, \tag{2.19}$$

$$\|[\widehat{\Delta A}, \widehat{\Delta b}\,\theta]\|_F^2 = (1-\rho)^2\|ry^{\dagger}\|_2^2 + \rho^2\theta^2\|r\|_2^2 = \rho\theta^2\|r\|_2^2. \tag{2.20}$$

In Case 2 we will show that if $\lambda_{\min}(M) \geq 0$ then $w = 0$ is optimal.

Case 2: The optimal $w \neq 0$. Let Y be as in Case 1, and $W = [w/\|w\|_2, W_2] \in \mathbb{R}^{m\times m}$ be an orthogonal matrix. Since $w^T(b + \Delta b) = 0$ we can write

$$b + \Delta b = W_2 s \quad \text{ for some } s \in \mathbb{R}^{m-1}. \tag{2.21}$$

From (2.10) we have

$$\begin{aligned} W^T\Delta A Y &= \begin{bmatrix} w^T/\|w\|_2 \\ W_2^T \end{bmatrix} \left[(b + \Delta b - w)y^{\dagger} + (I - ww^{\dagger})Z(I - yy^{\dagger}) - A\right][y/\|y\|_2, Y_2] \\ &= \left[\begin{array}{c|c} -\|w\|_2/\|y\|_2 - w^TAy/(\|w\|_2\|y\|_2) & -w^TAY_2/\|w\|_2 \\ \hline s/\|y\|_2 - W_2^TAy/\|y\|_2 & W_2^TZY_2 - W_2^TAY_2 \end{array}\right]. \end{aligned}$$

Thus the objective function can be written as five additive nonnegative terms:

$$\begin{aligned} \|[\Delta A, \Delta b\,\theta]\|_F^2 = &\left[\|w\|_2/\|y\|_2 + w^TAy/(\|w\|_2\|y\|_2)\right]^2 + \|w^TAY_2\|_2^2/\|w\|_2^2 \\ &+ \|s - W_2^TAy\|_2^2/\|y\|_2^2 + \|W_2^T(Z - A)Y_2\|_F^2 + \theta^2\|W_2 s - b\|_2^2. \end{aligned} \tag{2.22}$$

To minimize this we take $Z = \widehat{Z} \equiv A$ and note the sum of terms involving s is

$$\begin{aligned} \phi(s) &\equiv \|s - W_2^TAy\|_2^2/\|y\|_2^2 + \theta^2\|W_2 s - b\|_2^2 \\ &= \left[\|W^T(W_2 s - Ay)\|_2^2 - (w^TAy)^2/\|w\|_2^2 + \theta^2\|y\|_2^2\|W_2 s - b\|_2^2\right]/\|y\|_2^2 \\ &= \frac{1}{\|y\|_2^2}\left\|\begin{bmatrix} I \\ \theta\|y\|_2 I \end{bmatrix} W_2 s - \begin{bmatrix} Ay \\ b\,\theta\|y\|_2 \end{bmatrix}\right\|_2^2 - \left(\frac{w^TAy}{\|w\|_2\|y\|_2}\right)^2. \end{aligned} \tag{2.23}$$

The normal equations for $\hat{s}$, the optimal s, give $(1 + \theta^2\|y\|_2^2)\hat{s} = W_2^T(Ay + b\,\theta^2\|y\|_2^2)$. Therefore

$$\hat{s} = W_2^T[Ay\rho + b(1-\rho)] = W_2^T(b - r\rho). \tag{2.24}$$

Substituting this in the first line of (2.23) gives with $W_2W_2^T = I - ww^{\dagger}$

$$\begin{aligned} \phi(\hat{s}) &= \|W_2^Tr(1-\rho)\|_2^2/\|y\|_2^2 + \theta^2\|ww^{\dagger}b + W_2W_2^Tr\rho\|_2^2 \\ &= \rho\theta^2\|(I - ww^{\dagger})r\|_2^2 + \theta^2\|ww^{\dagger}b\|_2^2. \end{aligned}$$

Then from (2.22), we obtain

$$\min_{[\Delta A, \Delta b] \in \mathcal{C}^+_{A,b}} \|[\Delta A, \Delta b\,\theta]\|_F^2 = \min_w [\psi_1(w) + \psi_2(w)], \tag{2.25}$$

$$\psi_1(w) \equiv \left[\|w\|_2/\|y\|_2 + w^TAy/(\|w\|_2\|y\|_2)\right]^2, \tag{2.26}$$

$$\psi_2(w) \equiv \|w^TAY_2\|_2^2/\|w\|_2^2 + \rho\theta^2\|(I - ww^{\dagger})r\|_2^2 + \theta^2\|ww^{\dagger}b\|_2^2. \tag{2.27}$$

We will minimize $\psi_2(w)$, which is a function of $w/\|w\|_2$ alone, and then show that we can set $\psi_1(w)$ to zero by scaling w, leading to the optimal w. Since $Y_2Y_2^T = I - yy^\dagger$,

$$\psi_2(w) = \frac{w^TA(I-yy^\dagger)A^Tw}{w^Tw} + \rho\theta^2\|r\|_2^2\frac{w^T(I-rr^\dagger)w}{w^Tw} + \theta^2\frac{w^Tbb^Tw}{w^Tw} \tag{2.28}$$
$$= \frac{w^TNN^Tw}{w^Tw} = \rho\theta^2\|r\|_2^2 + \frac{w^TMw}{w^Tw},$$

whose minimum is $\rho\theta^2\|r\|_2^2 + \lambda_{\min}(M)$ given by $w = w_\theta\alpha$ for any nonzero $\alpha \in \mathbb{R}$ and w_θ satisfying $Mw_\theta = w_\theta\lambda_{\min}(M)$, $\|w_\theta\|_2 = 1$, since we assumed $w \neq 0$.

If $\lambda_{\min}(M) \geq 0$ the above with (2.20) in Case 1 show that $w = 0$ is optimal for minimizing $\|[\Delta A, \Delta b\,\theta]\|_F^2$, giving the minimum value $\rho\theta^2\|r\|_2^2$. So from (2.18), (2.19) and (2.20) we see that the top equalities in each of (2.15), (2.16) and (2.17) hold. Only when $\lambda_{\min}(M) < 0$ do we need to consider the possibility that $w \neq 0$.

Assume that $\lambda_{\min}(M) < 0$. It is easy to verify that $\psi_1(\hat{w}) = 0$ if

$$\hat{w} \equiv -w_\theta(w_\theta^TAy) \neq 0. \tag{2.29}$$

Suppose $w_\theta^TAy = 0$, then from $w_\theta^TMw_\theta = \lambda_{\min}(M) < 0$ and (2.14)

$$0 > \lambda_{\min}(M) = w_\theta^TAA^Tw_\theta - (w_\theta^Tr)^2\rho\theta^2 + (w_\theta^Tb)^2\theta^2 = w_\theta^TAA^Tw_\theta + (w_\theta^Tb)^2(1-\rho)\theta^2$$

which is impossible since the right hand side is nonnegative, see (2.12), proving that the inequality in (2.29) holds. Therefore from (2.25) we see that when $\lambda_{\min}(M) < 0$ the extended minimal backward error $\mu_F(y,\theta)$ satisfies the two bottom equalities in (2.15). The bottom equality in (2.17) follows immediately from (2.21) and (2.24), and substituting this with $Z = A$ and (2.29) in (2.10) gives

$$\begin{aligned} A + \Delta A &= [(I-w_\theta w_\theta^\dagger)(b-r\rho) + w_\theta(w_\theta^TAy)]y^\dagger + (I-w_\theta w_\theta^\dagger)A(I-yy^\dagger) \\ &= (I-w_\theta w_\theta^\dagger)[A + (b-r\rho-Ay)y^\dagger] + w_\theta(w_\theta^TAy)y^\dagger \\ &= (I-w_\theta w_\theta^\dagger)[A + r(1-\rho)y^\dagger] + w_\theta(w_\theta^TAy)y^\dagger, \end{aligned}$$

to prove the bottom equation in (2.16). □

REMARK 2.1. *The criterion $\lambda_{\min}(M(\theta)) \geq 0$ appears in (2.15)–(2.17). But if, as is usual, $m > n+1$, then $M(\theta)$ has at least $m-n-1$ zero eigenvalues corresponding to eigenvectors spanning $\mathcal{R}([A,b])^\perp$. Eigenvalues of a parameterized matrix that are zero independent of the parameter (here θ) will be called "trivial zero eigenvalues". Because they remain zero, their limiting behavior is trivial.*

2.2. Allowing a backward perturbation in A alone. In DLS problems only the matrix A is assumed to have uncertainty, so it is also important to consider the case where there is a backward perturbation in A alone. Then the corresponding minimal backward error problem becomes

$$\min_{\Delta A \in \mathcal{C}_A} \|\Delta A\|_F \qquad \text{where}$$

$$\mathcal{C}_A^+ \equiv \{\Delta A : (A+\Delta A)^T[b-(A+\Delta A)y] = -y\frac{\|b-(A+\Delta A)y\|_2^2}{\|y\|_2^2}\}, \tag{2.30}$$

$$\mathcal{C}_A \equiv \big\{\Delta A :\ \Delta A \in \mathcal{C}_A^+ \ \&\ \frac{\|b-(A+\Delta A)y\|_2}{\|y\|_2} < \sigma_{min}(A+\Delta A)\big\}, \tag{2.31}$$

and these two sets are just $\mathcal{C}^+_{A,b}$ and $\mathcal{C}_{A,b}$ in (2.8) and (2.9) with $\Delta b = 0$, so that $\Delta A \in \mathcal{C}^+_A \Rightarrow [\Delta A, 0] \in \mathcal{C}^+_{A,b}$. We can force $\Delta b = 0$ by taking the limit as $\theta \to \infty$ in (2.11), giving for the *extended* minimal backward error in this more limited case

$$\mu_F(y) \equiv \min_{\Delta A \in \mathcal{C}^+_A} \|\Delta A\|_F = \lim_{\theta\to\infty} \mu_F(y,\theta). \tag{2.32}$$

Here we have abused the notation a little by using both $\mu_F(\cdot)$ and $\mu_F(\cdot,\cdot)$. The proof using $\theta\to\infty$ (we use $\varepsilon \equiv \theta^{-2} \searrow 0$) is made possible by a beautiful classical result.

LEMMA 2.3. *(Rellich [24, pp. 29–37], see also Kato [16, pp. 121–2]). For $\varepsilon \in \mathbb{R}$ suppose $H(\varepsilon) = H(\varepsilon)^T \in \mathbb{R}^{n\times n}$ is analytic about $\varepsilon = 0$, then its eigenvalues and an orthonormal set of eigenvectors can be chosen analytic about $\varepsilon = 0$.*

We need some more results to prepare for Theorem 2.6.

LEMMA 2.4. *If $M = SWS^T$ with $S \in \mathbb{R}^{m\times n}$, $W = W^T \in \mathbb{R}^{n\times n}$, $m \ge n$, then*

(a) M has no more positive (negative) eigenvalues than W.

(b) If $W = I - y\alpha y^\dagger$ then M has at most one negative eigenvalue.

Proof. Part (a) was proven in [15, §4.5.11] for the case $m = n$. For $m > n$ writing $M = [S,0]\,\mathrm{diag}(W,0)[S,0]^T$ with square $[S,0]$ proves that (a) still holds. Since $I - y\alpha y^\dagger$ has eigenvalues 1 when $y = 0$, and $1-\alpha, 1, \ldots, 1$ when $y \ne 0$, (b) follows from (a). ∎

To make later analysis easier, we use ε to replace θ^{-2}. From (2.12)

$$\rho = \varepsilon/(\varepsilon + \|y\|_2^2), \qquad \varepsilon + \rho\|y\|_2^2 = \varepsilon(2-\rho), \qquad \varepsilon \equiv \theta^{-2}. \tag{2.33}$$

In our limits we only consider $\varepsilon \ge 0$, so $\varepsilon \to 0$ will always mean $\varepsilon \searrow 0$.

THEOREM 2.5. *With (2.33) if $A \in \mathbb{R}^{m\times n}$, $0 \ne y \in \mathbb{R}^n$, $0 \ne b \in \mathbb{R}^m$, $r \equiv b - Ay$,*

$$H(\varepsilon) \equiv \varepsilon A(I - yy^\dagger)A^T - r\rho r^T + bb^T,\ \varepsilon \in \mathbb{R},\ H(\varepsilon)w(\varepsilon) = w(\varepsilon)\lambda(\varepsilon),\ w(\varepsilon) \in \mathbb{R}^m, \tag{2.34}$$

$m \ge 2$ and $\lambda(0) \equiv \lambda_{\min}(H(0))$, then for small enough $\varepsilon \ge 0$ the minimum eigenvalue $\lambda(\varepsilon)$ and its normalized eigenvector $w(\varepsilon)$ can be chosen analytic with the forms

$$\lambda(\varepsilon) = \lambda_1\varepsilon + \lambda_2\varepsilon^2 + \cdots;\ w(\varepsilon) = w_0 + w_1\varepsilon + w_2\varepsilon^2 + \cdots,\ b^T w_0 = 0,\ \|w(\varepsilon)\|_2 = 1. \tag{2.35}$$

$$\text{If } T(\varepsilon) \equiv \varepsilon^{-1} P_b^\perp H(\varepsilon) P_b^\perp, \quad \text{then } T(0) \equiv \lim_{\varepsilon\to 0} T(\varepsilon) = P_b^\perp A(I - y2y^\dagger)A^T P_b^\perp. \tag{2.36}$$

Let $\lambda_*(0) \equiv \lambda_{\min}(T(0))$, then for small enough $\varepsilon \ge 0$ the minimum eigenvalue $\lambda_*(\varepsilon)$ and its normalized eigenvector $w_*(\varepsilon)$ can be chosen analytic in

$$T(\varepsilon)w_*(\varepsilon) = w_*(\varepsilon)\lambda_*(\varepsilon), \qquad w_*(\varepsilon) \in \mathbb{R}^m, \quad \|w_*(\varepsilon)\|_2 = 1. \tag{2.37}$$

Finally $\lim_{\varepsilon\to 0} \varepsilon^{-1}\lambda(\varepsilon)$ exists (see (2.35)), and for $\lambda_*(0)$ and $w_*(0)$ in (2.37),

$$\lim_{\varepsilon\to 0} b^T w(\varepsilon) = 0, \qquad \lim_{\varepsilon\to 0} \varepsilon^{-\frac{1}{2}} b^T w(\varepsilon) = 0, \qquad \lim_{\varepsilon\to 0} \varepsilon^{-1} b^T w(\varepsilon) = b^T w_1; \tag{2.38}$$

$$\lim_{\varepsilon\to 0} \varepsilon^{-1}\lambda(\varepsilon) < 0 \ \Rightarrow\ \lambda_*(0) < 0 \ \Rightarrow\ \lim_{\varepsilon\to 0} \varepsilon^{-1}\lambda(\varepsilon) = \lambda_*(0) \ \ \&\ \ w(0) = \pm w_*(0); \tag{2.39}$$

$$\lim_{\varepsilon\to 0} \varepsilon^{-1}\lambda(\varepsilon) \ge 0 \ \Leftrightarrow\ \lambda_*(0) = 0; \tag{2.40}$$

$$\lim_{\varepsilon\to 0} \varepsilon^{-1}\lambda(\varepsilon) = 0 \Rightarrow \lambda_*(0) = 0\ \&\ \begin{cases} \{\exists x \text{ such that } b = Ax\ \&\ \angle(x,y) = \pm\pi/4\}, \\ \text{or } \{\varepsilon^{-1}\lambda(\varepsilon) = 0 \text{ in a neighborhood of } \varepsilon = 0\}. \end{cases} \tag{2.41}$$

Proof. The expression for $T(0)$ in (2.36) follows from (2.34) and (2.33), and then

$$T(\varepsilon) = T(0) + P_b^\perp A y\rho y^\dagger A^T P_b^\perp = P_b^\perp A[I - y(2-\rho)y^\dagger]A^T P_b^\perp. \tag{2.42}$$

Clearly $H(\varepsilon)$ and $T(\varepsilon)$ are analytic about $\varepsilon=0$, so $w(\varepsilon)$, $\lambda(\varepsilon)$, $w_*(\varepsilon)$, and $\lambda_*(\varepsilon)$ can be chosen to be analytic with $\|w(\varepsilon)\|_2 = \|w_*(\varepsilon)\|_2 = 1$, see Lemma 2.3. Also $H(\varepsilon)$ can have at most one negative eigenvalue, see the start of the proof of Theorem 2.2, so from Lemma 2.4 $T(\varepsilon)$ can have at most one negative eigenvalue. Since $m \geq 2$, $H(0)=bb^T$ has minimum eigenvalue $\lambda(0)=0$, proving the first part of (2.35). Since $bb^Tw(0)=w(0)\lambda(0)=0$, we must have $b^Tw(0)=b^Tw_0=0$, proving the rest of (2.35). Next (2.35) proves (2.38), and we have

$$\begin{aligned}(2.43)\quad \lim_{\varepsilon\to 0}\varepsilon^{-1}\lambda(\varepsilon) &= \lim_{\varepsilon\to 0}\varepsilon^{-1}w(\varepsilon)^TH(\varepsilon)w(\varepsilon) = \lim_{\varepsilon\to 0}\varepsilon^{-\frac{1}{2}}w(\varepsilon)^TP_b^{\perp}H(\varepsilon)P_b^{\perp}w(\varepsilon)\varepsilon^{-\frac{1}{2}}\\ &= \lim_{\varepsilon\to 0} w(\varepsilon)^TT(\varepsilon)w(\varepsilon) = w(0)^TT(0)w(0) \geq \lambda_*(0) = \lambda_{\min}(T(0)),\end{aligned}$$

so $\lim_{\varepsilon\to 0}\varepsilon^{-1}\lambda(\varepsilon)<0 \Rightarrow \lambda_*(0)<0$. When $\lambda_*(0)<0$ it is a singleton, see Lemma 2.4, so for small enough ε, $\lambda_*(\varepsilon)<0$; then $b^Tw_*(\varepsilon)=0$ from (2.37) and (2.36), giving

$$(2.44)\qquad \lambda_*(\varepsilon) = \varepsilon^{-1}w_*(\varepsilon)^TP_b^{\perp}H(\varepsilon)P_b^{\perp}w_*(\varepsilon) = \varepsilon^{-1}w_*(\varepsilon)^TH(\varepsilon)w_*(\varepsilon).$$

Taking the limit as $\varepsilon\to 0$ and using (2.43) with $\|w(\varepsilon)\|_2=\|w_*(\varepsilon)\|_2=1$ gives

$$\begin{aligned}\lambda_*(0) &= \lim_{\varepsilon\to 0}\varepsilon^{-1}w_*(\varepsilon)^TH(\varepsilon)w_*(\varepsilon) \geq \lim_{\varepsilon\to 0}\varepsilon^{-1}\lambda(\varepsilon) = w(0)^TT(0)w(0)\\ &\geq \lambda_*(0) = w_*(0)^TT(0)w_*(0),\end{aligned}$$

proving equality throughout, so that $\lim_{\varepsilon\to 0}\varepsilon^{-1}\lambda(\varepsilon)=\lambda_*(0)$ when $\lambda_*(0)<0$. Also $w(0)^TT(0)w(0)=w_*(0)^TT(0)w_*(0)$ is a minimum of $w^TT(0)w$ over $w^Tw=1$ with unique minimizer (up to sign) when $\lambda_*(0)$ is a singleton, completing the proof of (2.39). Since $T(\varepsilon)b=0$ we see that $\lambda_*(0)\leq 0$, and (2.40) follows using (2.39).

Now assume $\lim_{\varepsilon\to 0}\varepsilon^{-1}\lambda(\varepsilon)=0$, ($\lambda_1=0$ in (2.35)). Then $\lambda_*(0)=0$ in (2.41) follows from (2.40). If $[A,b]$ has rank s then $\varepsilon^{-1}H(\varepsilon)$ has m–s trivial zero eigenvalues. If in the limit as $\varepsilon\to 0$ there are only trivial zero eigenvalues then by continuity $\varepsilon^{-1}\lambda(\varepsilon)=0$ in some neighborhood of $\varepsilon=0$. Next assume there is a nontrivial zero eigenvalue, that is, an eigenpair of the form (2.35) with $\lambda_1=0$ and $A^Tw_0\neq 0$. But $\lambda_*(0)=0$ shows that $T(0)$ is positive semi-definite, and $0=\lim_{\varepsilon\to 0}\varepsilon^{-1}\lambda(\varepsilon)=\lim_{\varepsilon\to 0}w_0^T\varepsilon^{-1}H(\varepsilon)w_0=\lim_{\varepsilon\to 0}w_0^TT(\varepsilon)w_0$, see (2.36), so $0=w_0^TT(0)w_0=w_0^TA(I-y2y^\dagger)A^Tw_0$. Here $I-y2y^\dagger$ is a Householder reflection. Thus

$$(2.45)\qquad \|A^Tw_0\|_2^2 = 2(w_0^TAy)^2/\|y\|_2^2, \qquad 0=T(0)w_0=P_b^{\perp}A(I-y2y^\dagger)A^Tw_0,$$

and $A(I-y2y^\dagger)A^Tw_0 = bb^\dagger A(I-y2y^\dagger)A^Tw_0\neq 0$, see (2.2). Then $Ax\nu=b\nu$ where

$$x\nu = (I-y2y^\dagger)A^Tw_0, \quad \nu\equiv b^TA(I-y2y^\dagger)A^Tw_0/b^Tb\neq 0.$$

From this $y^Tx\nu=-y^TA^Tw_0$, and with (2.45) $x^Tx\nu^2=\|A^Tw_0\|_2^2=2(w_0^TAy)^2/\|y\|_2^2=2(y^Tx)^2\nu^2/\|y\|_2^2$. This gives $2(y^Tx)^2=y^Ty\cdot x^Tx$, so $\angle(x,y)=\pm\pi/4$, proving (2.41). ∎

REMARK 2.2. *In (2.41) the case $b=Ax$ & $\angle(x,y)=\pm\pi/4$ is extremely unlikely (it has "probability zero"), requiring $b\in\mathcal{R}(A)$ (a highly unlikely situation when we are solving DLS problems), and y to be a terrible approximation to the unique x for which $b=Ax$, giving* exactly $\angle(x,y)=\pm\pi/4$.

We can now obtain the equivalent of Theorem 2.2 for a backward perturbation restricted to A alone.

THEOREM 2.6. *Suppose that we are given $A \in \mathbb{R}^{m\times n}$, $m \geq 2$, with nonzero $b \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$; suppose also that (2.2) holds. Let $r \equiv b - Ay$ and*

$$N_\infty \equiv \Big[(I-bb^\dagger)A(I-yy^\dagger), \frac{\|r\|_2}{\|y\|_2}(I-bb^\dagger)(I-rr^\dagger), \frac{b}{\|b\|_2}\frac{\|r\|_2}{\|y\|_2}\Big], \tag{2.46}$$

$$M_\infty = M_\infty(y) \equiv (I-bb^\dagger)A\left(I-y2y^\dagger\right)A^T(I-bb^\dagger) = N_\infty N_\infty^T - \|r\|_2^2/\|y\|_2^2\, I. \tag{2.47}$$

Then $\lambda_{\min}(M_\infty) \leq 0$ and M_∞ has at most one negative eigenvalue. Also the DLS extended minimal backward error $\mu_F(y)$ in (2.32) satisfies

$$\mu_F^2(y) = \frac{\|r\|_2^2}{\|y\|_2^2} + \lambda_{\min}(M_\infty) = \sigma_{\min}^2(N_\infty). \tag{2.48}$$

Furthermore $\mu_F(y)$ is given by the backward perturbation:

$$\widehat{\Delta A} = \begin{cases} ry^\dagger & \text{if } \lambda_{\min}(M_\infty) = 0, \\ ry^\dagger - w_* w_*^\dagger A(I - y2y^\dagger) & \text{if } \lambda_{\min}(M_\infty) < 0, \end{cases} \tag{2.49}$$

where w_* is the eigenvector of M_∞ corresponding to $\lambda_{\min}(M_\infty) < 0$, or equivalently the left singular vector of N_∞ corresponding to $\sigma_{\min}(N_\infty)$. If $\lambda_{\min}(M_\infty) < 0$ in (2.47), then $\lambda_{\min}(M_\infty) = \lim_{\theta\to\infty}\lambda_{\min}(M(\theta))$ in (2.14). In general (2.48)–(2.49) are the corresponding limiting values of (2.15)–(2.16) as $\theta \to \infty$, except possibly in the probability zero case mentioned in Remark 2.2 that could only occur if $\lim_{\theta\to\infty}\lambda_{\min}(M(\theta)) = 0$ is a nontrivial zero eigenvalue, see Remark 2.1.

Proof. Since (2.47) follows from (2.46), the results on N_∞ follow trivially from those on M_∞. Since $M_\infty b = 0$, $\lambda_{\min}(M_\infty) \leq 0$. From Lemma 2.4 M_∞ has at most one negative eigenvalue. In fact M_∞ in (2.47) is identical to $T(0)$ in (2.36), so with $w_* \equiv w_*(0)$, $\lambda_* \equiv \lambda_{\min}(M_\infty) \equiv \lambda_*(0)$ in (2.37), $M_\infty w_* = w_*\lambda_*$. From (2.14) and (2.34)

$$M(\theta) \equiv A\left(I-yy^\dagger\right)A^T - r\rho\theta^2 r^T + b\,\theta^2 b^T = \varepsilon^{-1}H(\varepsilon) \;\text{ with } \varepsilon \equiv \theta^{-2}. \tag{2.50}$$

Since $M(\theta)$ can have at most one negative eigenvalue, when $\lim_{\theta\to\infty}\lambda_{\min}(M(\theta)) < 0$, $\lambda_{\min}(M(\theta))$ is the unique minimum eigenvalue for large enough θ, and w_θ in (2.16) can be taken as $w(\varepsilon)$ in Theorem 2.5. This, and noting that $\lambda(\theta^2 H(\theta^{-2}))$ is equal to $\lambda_{\min}(M(\theta))$ for large enough θ, gives from (2.38)–(2.41)

$$\lim_{\theta\to\infty}\lambda_{\min}(M(\theta)) < 0 \;\Rightarrow\; \begin{cases} \lambda_{\min}(M_\infty) = \lim_{\theta\to\infty}\lambda_{\min}(M(\theta)) \\ \;\&\;\; w_* = \pm\lim_{\theta\to\infty} w_\theta \;\;\&\;\; \lim_{\theta\to\infty}\theta w_\theta^T b = 0; \end{cases}$$

$$\lim_{\theta\to\infty}\lambda_{\min}(M(\theta)) \geq 0 \;\Leftrightarrow\; \lambda_* \equiv \lambda_{\min}(M_\infty) = 0;$$

$$\lim_{\theta\to\infty}\lambda_{\min}(M(\theta)) = 0 \;\Rightarrow\; \lambda_{\min}(M_\infty) = 0 \;\&\; \begin{cases} \{\exists x \;:\; b = Ax \;\&\; \angle(x,y) = \pm\pi/4\}, \\ \text{or } \{\exists\theta_1 \;:\; \lambda_{\min}(M(\theta)) = 0 \;\forall\theta > \theta_1\}. \end{cases}$$

But from (2.12) $\lim_{\theta\to\infty}\rho = 0$ and $\lim_{\theta\to\infty}\rho\theta^2 = \|y\|_2^{-2}$, so that (2.48) is the limiting value of both cases of (2.15) as $\theta \to \infty$. If $\lim_{\theta\to\infty}\lambda_{\min}(M(\theta)) \neq 0$ it can also be seen that in the limit the two criteria in (2.16) become the respective criteria in (2.49), where if $\lim_{\theta\to\infty}\lambda_{\min}(M(\theta)) < 0$, $b^T w_* = \lim_{\theta\to\infty} b^T w_\theta = 0$, so that in the limit the two expressions for $\widehat{\Delta A}$ in (2.16) become the respective expressions in (2.49). If $\lim_{\theta\to\infty}\lambda_{\min}(M(\theta)) = 0$ and this corresponds to trivial zero eigenvalues only, the top row of (2.49) is clearly once again the correct limit. Only the probability zero case of $\lim_{\theta\to\infty}\lambda_{\min}(M(\theta)) = 0$ with $b = Ax$, $\angle(x,y) = \pm\pi/4$, allows the possibility

that $\lambda_{\min}(M(\theta)) < 0$ for arbitrarily large θ, and since in this case $\lambda_{\min}(M_\infty)=0$, this suggests that (2.16) could fail to give the correct limiting perturbation in (2.49). That is, although (2.49) is correct, until proven otherwise there remains the possibility that taking the limit in (2.16) could lead to $\widehat{\Delta A} = ry^\dagger - w_* w_*^\dagger A(I - y2y^\dagger)$ rather than $ry^\dagger$ in this one strange case, see (2.49). □

Deriving $\mu_F(y)$ directly as we did for $\mu_F(y,\theta)$ also leads to the results down to the sentence including (2.49). But Theorem 2.5 describes the limiting behavior as well.

To parallel the remark given in [13, Sec. 20.7] for some formulas for the minimal backward error of ordinary least squares problems, computing and adding the eigenvalue in (2.15) or (2.48) is not wise computationally. Catastrophic cancellation may occur when it is negative. Furthermore the computed value may have very poor accuracy even using well-known software such as MATLAB 7.0, e.g., in (2.48) the computed value of $\lambda_{\min}(M_\infty)$ may be smaller than $-\|r\|_2^2/\|y\|_2^2$. The singular value is much more reliable for computation. If we computed that using the Golub-Reinsch singular value decomposition algorithm it would need about $8/3m^3 + 4mn^2$ flops, but one point of this paper is that we can use cheaper lower bounds or estimates instead.

In Theorem 2.2 we have either $\lambda_{\min}(M(\theta)) < 0$ or $\lambda_{\min}(M(\theta)) \geq 0$, while in Theorem 2.6 we have either $\lambda_{\min}(M_\infty) < 0$ or $\lambda_{\min}(M_\infty) = 0$. By substituting the resulting perturbations in the relevant inequalities it is straightforward to see that the inequality in (2.9) is satisfied when $\lambda_{\min}(M(\theta)) \geq 0$ and $\operatorname{rank}(A + r(1-\rho)y^\dagger) = n$ and the inequality in (2.31) is satisfied when $\lambda_{\min}(M_\infty) = 0$ and $\operatorname{rank}(A + ry^\dagger) = n$. It follows that in these two special cases the extended minimal backward error is actually the true minimal backward error, and that nothing was lost by using the "supersets" $\mathcal{C}_{A,b}^+$ and $\mathcal{C}_A^+$. We will supply further justification for the use of these supersets later.

The following result indicates that the extended minimal backward error $\mu_F(y)$ is continuous at $y = \hat{x}$, where $\hat{x}$ is the DLS solution in (2.5)–(2.6), where of course $\mu_F(\hat{x}) = 0$. In order to save space, here and in the rest of the paper we will only consider the case where A is perturbed, but all the results could be extended to the more general case where both A and b are perturbed.

COROLLARY 2.7. *With the notation and conditions of Theorem 2.6, and the DLS solution $\hat{x}$ in (2.5)–(2.6), define $\hat{M}_\infty \equiv M_\infty(\hat{x})$ (see (2.47)), and $\hat{r} \equiv b - A\hat{x}$. Then with (2.15),*

$$\lim_{y \to \hat{x}} \mu_F(y) = \mu_F(\hat{x}) = \left(\frac{\|\hat{r}\|_2^2}{\|\hat{x}\|_2^2} + \lambda_{\min}(\hat{M}_\infty) \right)^{1/2} = 0.$$

Proof. First we see that (2.5) just says

$$A^T\hat{r} = A^T(b - A\hat{x}) = -\hat{x}\|b - A\hat{x}\|_2^2/\|\hat{x}\|_2^2 = -\hat{x}\|\hat{r}\|_2^2/\|\hat{x}\|_2^2,$$

and multiplying this on the left by $\hat{x}^T$ shows that

$$0 = (b - A\hat{x})^T(b - A\hat{x}) + (A\hat{x})^T(b - A\hat{x}) = b^T(b - A\hat{x}) = b^T\hat{r}, \tag{2.51}$$

so that $(I - bb^\dagger)\hat{r} = \hat{r}$. Since $\hat{M}_\infty = M_\infty(\hat{x}) = (I - bb^\dagger)A(I - 2\hat{x}\hat{x}^\dagger)A^T(I - bb^\dagger)$, the above give

$$\begin{aligned}
\hat{M}_\infty\hat{r} &= (I - bb^\dagger)A(I - 2\hat{x}\hat{x}^\dagger)A^T\hat{r} = -(I - bb^\dagger)A(I - 2\hat{x}\hat{x}^\dagger)\hat{x}\|\hat{r}\|_2^2/\|\hat{x}\|_2^2 \\
&= (I - bb^\dagger)A\hat{x}\|\hat{r}\|_2^2/\|\hat{x}\|_2^2 = (I - bb^\dagger)(A\hat{x} - b)\|\hat{r}\|_2^2/\|\hat{x}\|_2^2 \\
&= -\hat{r}\|\hat{r}\|_2^2/\|\hat{x}\|_2^2.
\end{aligned}$$

Thus by Lemma 2.4, $-\|\hat{r}\|_2^2/\|\hat{x}\|_2^2$ is the only negative eigenvalue of $\hat{M}_\infty$, and $\mu_F(\hat{x}) = 0$. Clearly when $y \to \hat{x}$ we have $r = b - Ay \to \hat{r}$, $M_\infty \to \hat{M}_\infty$, and by the continuity of the eigenvalues of M_∞ in (2.47), $\lambda_{\min}(M_\infty) \to \lambda_{\min}(\hat{M}_\infty)$, completing the proof. ∎

Since in (2.32) $\mathcal{C}_A \subseteq \mathcal{C}_A^+$,

$$\mu_F(y) \equiv \min_{\Delta A \in \mathcal{C}_A^+} \|\Delta A\|_F \le \min_{\Delta A \in \mathcal{C}_A} \|\Delta A\|_F,$$

i.e., $\mu_F(y)$ is a lower bound on the minimal backward error. However we have found computationally, see section 5, that when y is a reasonable approximation to the exact solution of the DLS problem (2.1), the minimal perturbation $\widehat{\Delta A}$ usually satisfies the inequality in (2.31). Therefore in such cases $\mu(y)$ is actually the minimal backward error. The following result partially justifies this finding.

THEOREM 2.8. *For given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ suppose that (2.2) holds. Let $\hat{x}$ be the DLS solution to* (2.1)*. Then there exists an $\epsilon > 0$ such that if $\|y - \hat{x}\|_2 < \epsilon$, then $\mu_F(y)$ above is the true minimal backward error.*

Proof. For any given y Theorem 2.6 shows that $\widehat{\Delta A}$ satisfying (2.49) is the minimizer of (2.32). Notice that when $y \to \hat{x}$ we have from Corollary 2.7 that $\widehat{\Delta A} \to 0$. Thus

$$\lim_{y\to\hat{x}} \left(\frac{\|b - (A + \widehat{\Delta A})y\|_2}{\|y\|_2} - \sigma_{\min}(A + \widehat{\Delta A}) \right) = \frac{\|b - A\hat{x}\|_2}{\|\hat{x}\|_2} - \sigma_{\min}(A)$$

Since $\frac{\|b-A\hat{x}\|_2}{\|\hat{x}\|_2} - \sigma_{\min}(A) < 0$, there must exist $\epsilon > 0$ such that when $\|y - \hat{x}\|_2 < \epsilon$,

$$\frac{\|b - (A + \widehat{\Delta A})y\|_2}{\|y\|_2} - \sigma_{\min}(A + \widehat{\Delta A}) < 0.$$

Therefore $\widehat{\Delta A} \in \mathcal{C}_A$ and $\mu_F(y) = \min_{\Delta A \in \mathcal{C}_A} \|\Delta A\|_F$, i.e., when $\|y - \hat{x}\|_2 < \epsilon$, $\mu_F(y)$ is the true minimal backward error. ∎

3. A lower bound on $\min_{\Delta A \in \mathcal{C}_A^+} \|\Delta A\|_2$. Since computing $\mu_F(y)$ directly is expensive, in this section we suggest a good lower bound which can be estimated easily.

First we give the following result, which is analogous to Theorem 3.1 in [35] for ordinary least squares problems.

THEOREM 3.1. *With the notation and conditions of Theorem 2.6, and with $\mathcal{C}_A^+$ in (2.30) let $\mu_2(y) \equiv \min_{\Delta A \in \mathcal{C}_A^+} \|\Delta A\|_2$. Then for $\mu_F(y)$ in (2.32)*

$$\frac{1}{\sqrt{2}}\mu_F(y) \le \mu_2(y) \le \mu_F(y). \tag{3.1}$$

Proof. For any $\Delta A \in \mathcal{C}_A^+$ we see from Lemma 2.1 that there exist w satisfying $b^T w = 0$ and $Z \in \mathbb{R}^{m\times n}$ such that

$$\begin{aligned}
\Delta A &= (b - w)y^\dagger + (I - ww^\dagger)Z(I - yy^\dagger) - A \\
&= [(I - ww^\dagger)b - w]y^\dagger + (I - ww^\dagger)Z(I - yy^\dagger) \\
&\quad - (I - ww^\dagger)Ayy^\dagger - ww^\dagger Ayy^\dagger - (I - ww^\dagger)A(I - yy^\dagger) - ww^\dagger A(I - yy^\dagger) \\
&= [(I - ww^\dagger)(b - Ay) - w - ww^\dagger Ay]y^\dagger - ww^\dagger A(I - yy^\dagger) \\
&\quad + (I - ww^\dagger)(Z - A)(I - yy^\dagger).
\end{aligned}$$

Denote $\Delta A_1 = [(I - ww^\dagger)(b - Ay) - w - ww^\dagger Ay]y^\dagger$, $\Delta A_2 = -ww^\dagger A(I - yy^\dagger)$, and $\Delta A_3 = (I - ww^\dagger)(Z - A)(I - yy^\dagger)$. Since $\Delta A_1 \Delta A_2^T = 0$, $\Delta A_1 \Delta A_3^T = 0$ and $\Delta A_2^T \Delta A_3 = 0$ and ΔA_1 and ΔA_2 are rank 1 matrices,

$$\begin{aligned}
\|\Delta A\|_2^2 &= \|\Delta A_1 + \Delta A_2 + \Delta A_3\|_2^2 = \|(\Delta A_1 + \Delta A_2 + \Delta A_3)(\Delta A_1 + \Delta A_2 + \Delta A_3)^T\|_2 \\
&= \|\Delta A_1 \Delta A_1^T + (\Delta A_2 + \Delta A_3)(\Delta A_2 + \Delta A_3)^T\|_2 \\
&\geq \max\{\|\Delta A_1 \Delta A_1^T\|_2, \|(\Delta A_2 + \Delta A_3)(\Delta A_2 + \Delta A_3)^T\|_2\} \\
&\geq \frac{1}{2}(\|\Delta A_1\|_2^2 + \|\Delta A_2 + \Delta A_3\|_2^2) \\
&= \frac{1}{2}(\|\Delta A_1\|_2^2 + \|(\Delta A_2 + \Delta A_3)^T(\Delta A_2 + \Delta A_3)\|_2) \\
&= \frac{1}{2}(\|\Delta A_1\|_2^2 + \|\Delta A_2^T \Delta A_2 + \Delta A_3^T \Delta A_3\|_2) \geq \frac{1}{2}(\|\Delta A_1\|_2^2 + \|\Delta A_2\|_2^2) \\
&= \frac{1}{2}(\|\Delta A_1\|_F^2 + \|\Delta A_2\|_F^2) = \frac{1}{2}\|\Delta A_1 + \Delta A_2\|_F^2 \geq \frac{1}{2} \min_{\Delta A \in \mathcal{C}_A^+} \|\Delta A\|_F^2,
\end{aligned}$$

where the last inequality is due to the fact that $\Delta A_1 + \Delta A_2 \in \mathcal{C}_A^+$ (take $Z = A$). Therefore

$$\min_{\Delta A \in \mathcal{C}_A^+} \|\Delta A\|_2^2 \geq \frac{1}{2} \min_{\Delta A \in \mathcal{C}_A^+} \|\Delta A\|_F^2,$$

leading to the first inequality in (3.1). The second inequality in (3.1) is easy to prove. In fact, if $\widehat{\Delta A}$ is a minimal solution to (2.32), then

$$\mu_F(y) = \|\widehat{\Delta A}\|_F \geq \|\widehat{\Delta A}\|_2 \geq \mu_2(y).$$

□

Now we give a lower bound on $\mu_2(y) \equiv \min_{\Delta A \in \mathcal{C}_A^+} \|\Delta A\|_2$.

THEOREM 3.2. *With the notation and conditions of Theorem 2.6, and with* $\mu_2(y) \equiv \min_{\Delta A \in \mathcal{C}_A^+} \|\Delta A\|_2$ *for* $\mathcal{C}_A^+$ *in (2.30),*

$$\mu_2(y) \geq \mu_2^{\rm lb}(y) \equiv \frac{2\beta_0}{\beta_1 + \sqrt{\beta_1^2 + 4\beta_0}}, \tag{3.2}$$

where

$$\beta_0 \equiv \frac{\big\|(A^T r\|y\|_2^2 + y\|r\|_2^2)\big\|_2}{2\|y\|_2^3}, \quad \beta_1 \equiv \frac{\|y\|_2^3\|A\|_2 + 3\|y\|_2^2\|r\|_2}{2\|y\|_2^3}. \tag{3.3}$$

Proof. For any $\Delta A \in \mathcal{C}_A^+$, from (2.30) we obtain

$$(A + \Delta A)^T(r - \Delta A y)\|y\|_2^2 + y\|r - \Delta A y\|_2^2 = 0.$$

Thus we have

$$\begin{aligned}
A^T r\|y\|_2^2 + y\|r\|_2^2 = {} & A^T \Delta A y\|y\|_2^2 - \Delta A^T r\|y\|_2^2 + y2(r^T \Delta A y) \\
& + \Delta A^T \Delta A y\|y\|_2^2 - y\|\Delta A y\|_2^2.
\end{aligned}$$

Then taking the 2-norm on both sides of this equation, we obtain the inequality

$$\big\|(A^T r\|y\|_2^2 + y\|r\|_2^2)\big\|_2 \leq (\|y\|_2^3\|A\|_2 + 3\|y\|_2^2\|r\|_2)\|\Delta A\|_2 + 2\|y\|_2^3\|\Delta A\|_2^2,$$

that is, with (3.3), the quadratic inequality in terms of $\xi \equiv \|\Delta A\|_2$:

$$\beta_0 \le \beta_1 \xi + \xi^2.$$

Since ξ and β_1 are nonnegative, $\xi \ge \xi_+$, where ξ_+ is the positive root of $\beta_0 = \beta_1\xi + \xi^2$, so

$$\xi \ge \xi_+ = (\sqrt{\beta_1^2 + 4\beta_0} - \beta_1)/2 = 2\beta_0/(\sqrt{\beta_1^2 + 4\beta_0} + \beta_1),$$

giving (3.2). ☐

The lower bound in (3.2) can usually be evaluated in $\mathcal{O}(mn)$ flops, since $\|A\|_2$ can usually be estimated by a standard norm estimator in $O(mn)$ flops, see [13, Sec. 15.2]. In fact a good estimate of $\|A\|_2$ might already be available from whatever method is used for obtaining y, and the cost will essentially be the $4mn$ flops for computing $A^T(b - Ay)$.

Also $\mu_2^{\rm lb}(\hat{x}) = \mu_2(\hat{x}) = \mu_F(\hat{x}) = 0$ as desired, see (2.5), Corollary 2.7, and (3.1).

4. An asymptotic estimate for $\mu_F(y)$. Computing $\mu_F(y)$ directly is expensive and the lower bound (3.2) may not be very tight. In this section we would like to give an asymptotic estimate by following the general approach given in [9].

Let $f(A, y) \equiv (b - Ay)^T(b - Ay)y + (y^Ty)A^T(b - Ay) = \|r\|_2^2 y + \|y\|_2^2 A^T r$. Note that $f(A, \hat{x}) = 0$ (see (2.5)). The extended minimal backward perturbation ΔA is the matrix satisfying $f(A + \Delta A, y) = 0$ and $\mu_F(y) = \|\Delta A\|_F$. But by Taylor's expansion, for small enough $E \in \mathbb{R}^{m\times n}$,

$$f(A + E, y) \approx f(A, y) + \mathcal{J}_A f(A, y)\mathtt{vec}(E),$$

where $\mathcal{J}_A f(A, y) \in \mathbb{R}^{n\times mn}$ is the Jacobian matrix of f with respect to $\mathtt{vec}(A)$. Thus an approximation to ΔA is that E giving the minimum 2-norm solution to

$$f(A, y) + \mathcal{J}_A f(A, y)\mathtt{vec}(E) = 0, \tag{4.1}$$

that is, E such that

$$\mathtt{vec}(E) = -[\mathcal{J}_A f(A, y)]^\dagger f(A, y), \quad \tilde{\mu}_F(y) \equiv \|E\|_F = \|[\mathcal{J}_A f(A, y)]^\dagger f(A, y)\|_2. \tag{4.2}$$

THEOREM 4.1. *With the notation and conditions of Theorem 2.6, $\tilde{\mu}_F(y)$ in (4.2) is an asymptotic estimate of $\mu_F(y)$ in (2.32), i.e. for $\hat{x}$ solving (2.1),*

$$\lim_{y\to\hat{x}} \frac{\tilde{\mu}_F(y)}{\mu_F(y)} = 1.$$

Proof. By Taylor's expansion,

$$0 = f(A + \Delta A, y) = f(A, y) + \mathcal{J}_A f(A, y)\mathtt{vec}(\Delta A) + O(\|\Delta A\|_F^2). \tag{4.3}$$

Thus from (4.2),

$$\mathtt{vec}(E) = -[\mathcal{J}_A f(A, y)]^\dagger f(A, y) = [\mathcal{J}_A f(A, y)]^\dagger \mathcal{J}_A f(A, y)\mathtt{vec}(\Delta A) + O(\|\Delta A\|_F^2).$$

Taking the 2-norm and noticing $[\mathcal{J}_A f(A, y)]^\dagger \mathcal{J}_A f(A, y)$ is an orthogonal projection matrix, we obtain

$$\tilde{\mu}_F(y) \le \mu_F(y) + O(\|\Delta A\|_F^2),$$

which, with Corollary 2.7, leads to $\lim_{y\to\hat{x}} \tilde{\mu}_F(y)/\mu_F(y) \le 1$.

On the other hand, from (4.1) and (4.3) we can obtain

$$\mathcal{J}_A f(A,y)\mathtt{vec}(\Delta A) = \mathcal{J}_A f(A,y)[\mathtt{vec}(E) + O(\|\Delta A\|_F^2)].$$

Since ΔA is a matrix satisfying the above equality with minimum F-norm, we must have

$$\|\mathtt{vec}(\Delta A)\|_2 \le \|\mathtt{vec}(E) + O(\|\Delta A\|_F^2)\|_2 \le \|\mathtt{vec}(E)\|_2 + O(\|\Delta A\|_F^2)$$

which, with Corollary 2.7, leads to $\lim_{y\to\hat{x}} \tilde{\mu}_F(y)/\mu_F(y) \ge 1$, completing the proof. □

Theorem 4.1 is similar to [9, Cor 3.4], where a general minimal backward error problem was considered, and applying the corollary to our case will result in the asymptotic estimate $\|[\mathcal{J}_A f(A,\hat{x})]^\dagger f(A,y)\|_2$. We thank the referee who pointed out that a general version of Theorem 4.1 was given in [10] (with no formal proof).

In the following we will consider computing or estimating $\tilde{\mu}_F(y)$. First we would like to obtain an explicit expression for it. If $f = (f_i)$ and $g = (g_i)$ are column vectors, then we define the matrix $\partial f/\partial g^T \equiv (\partial f_i/\partial g_j)$. Write $m \times n$ $A = [a_1, \ldots, a_n]$, $y^T = (\eta_1, \ldots, \eta_n)$, then $\partial r/\partial a_j^T = \partial(b - Ay)/\partial a_j^T = -\eta_j I$, $\partial(r^T r)/\partial a_j^T = 2r^T \partial r/\partial a_j^T = -2\eta_j r^T$, and if $i \ne j$, $\partial(a_i^T r)/\partial a_j^T = -\eta_j a_i^T$, while $\partial(a_j^T r)/\partial a_j^T = r^T - \eta_j a_j^T$, from which we see that, with $\mathtt{vec}(A)^T = (a_1^T, \ldots, a_n^T)$ and

$$\begin{aligned}
&\mathcal{J}_A f(A,y) \equiv \partial f(A,y)/\partial \mathtt{vec}(A)^T = [\partial f(A,y)/\partial a_1^T, \ldots, \partial f(A,y)/\partial a_n^T],\\
&\partial f(A,y)/\partial a_j^T = \partial(r^T r y + y^T y A^T r)/\partial a_j^T = -2\eta_j y r^T + y^T y e_j r^T - \eta_j y^T y A^T,\\
&\mathcal{J}_A f(A,y)\cdot[\mathcal{J}_A f(A,y)]^T = \sum_{j=1}^{n} [\partial f(A,y)/\partial a_j^T]\cdot[\partial f(A,y)/\partial a_j^T]^T\\
&= \sum_{j=1}^{n} (-2\eta_j y r^T + y^T y e_j r^T - \eta_j y^T y A^T)(-2\eta_j y r^T + y^T y e_j r^T - \eta_j y^T y A^T)^T\\
&= \|y\|_2^6 [A^T A + A^T r y^\dagger + (y^\dagger)^T r^T A + (\|r\|_2^2/\|y\|_2^2) I]\\
&= \|y\|_2^6 \begin{bmatrix} A + ry^\dagger \\ (\|r\|_2/\|y\|_2)(I - yy^\dagger) \end{bmatrix}^T \begin{bmatrix} A + ry^\dagger \\ (\|r\|_2/\|y\|_2)(I - yy^\dagger) \end{bmatrix}. \qquad (4.4)
\end{aligned}$$

Here the matrix $\begin{bmatrix} A + ry^\dagger \\ (\|r\|_2/\|y\|_2)(I - yy^\dagger) \end{bmatrix}$ has full column rank. In fact if it does not then there exists nonzero $x \in \mathbb{R}^n$ such that

$$(A + ry^\dagger)x = 0, \quad (I - yy^\dagger)x = 0,$$

and it follows that $Ay + r = 0$, so $b = 0$, contradicting our assumption (2.2). Therefore $\mathcal{J}_A f(A,y)$ has full row rank. Then from (4.2),

$$\begin{aligned}
\tilde{\mu}_F(y) &= \|[\mathcal{J}_A f(A,y]^T \left\{\mathcal{J}_A f(A,y)\cdot[\mathcal{J}_A f(A,y)]^T\right\}^{-1} f(A,y)\|_2\\
&= \|\left\{\mathcal{J}_A f(A,y)\cdot[\mathcal{J}_A f(A,y)]^T\right\}^{-1/2} f(A,y)\|_2, \qquad (4.5)
\end{aligned}$$

where the second equality can easily be proved by using the SVD of $\mathcal{J}_A f(A,y)$.

Define

$$B \equiv \begin{bmatrix} A + ry^\dagger \\ (\|r\|_2/\|y\|_2)(I - yy^\dagger) \end{bmatrix}, \quad c \equiv \begin{bmatrix} r \\ 0 \end{bmatrix} \in \mathbb{R}^{m+n}. \qquad (4.6)$$

Note that

$$f(A,y) = \|y\|_2^2(A^T r + \|r\|_2^2(y^\dagger)^T) = \|y\|_2^2 B^T c. \tag{4.7}$$

Then from (4.5) with (4.4) and (4.7), it follows that

$$\tilde{\mu}_F(y) = \frac{\|(B^TB)^{-1/2}B^Tc\|_2}{\|y\|_2} = \frac{[c^TB(B^TB)^{-1}B^Tc]^{1/2}}{\|y\|_2} = \frac{\|B(B^TB)^{-1}B^Tc\|_2}{\|y\|_2}. \tag{4.8}$$

Note that $B(B^TB)^{-1}B^T$ is an orthogonal projector onto $\mathcal{R}(B)$.

The asymptotic estimate $\tilde{\mu}_F(y)$ is analogous to an estimate for the minimal backward error for ordinary least squares problems whose various forms have been studied in [9], [11], [12], and [17]. One method for computing $\tilde{\mu}_F(y)$ is to use the QR factorization. If $B = QR$ where $Q \in \mathbb{R}^{(m+n)\times n}$ satisfies $Q^TQ = I_n$, and R is upper triangular, then we see that

$$\tilde{\mu}_F(y) = \|Q^Tc\|_2/\|y\|_2. \tag{4.9}$$

If we use Householder QR factorization, this method will cost $2(m + 2/3n^2)n^2$ flops. The other method is to use the moment method by following [27, Part I]. For brevity, we will not give details here.

5. Numerical tests. In section 2 we gave an extended minimal backward error $\mu_F(y)$, which is a lower bound on the minimal backward error. But if the inequality in (2.31) holds for the extended minimal backward perturbation $\widehat{\Delta A}$ given in (2.49), then $\mu_F(y)$ is in fact the minimal backward error. Our numerical tests indicate that if the given vector y is a reasonable approximation to the true DLS solution, the inequality in (2.31) holds, where ΔA is the minimal $\widehat{\Delta A}$ given in (2.16). We will give some examples in this section to illustrate this. In sections 3 we gave a lower bound $\mu_2^{\mathtt{lb}}(y)$ on $\mu_2(y)$, which is also a lower bound on $\mu_F(y)$ (since $\mu_2(y) \le \mu_F(y)$). In section 4 we presented an asymptotic estimate $\tilde{\mu}_F(y)$ of $\mu_F(y)$. We will give numerical examples to show how good $\mu_2^{\mathtt{lb}}(y)$ and $\tilde{\mu}_F(y)$ are as approximations to $\mu_F(y)$. We carried out computations using MATLAB 7.4 on a MacBook running Mac OS X 10.4.11.

In our numerical tests the data was constructed as follows (`randn` and `rand` are two MATLAB built-in functions for generating random matrices with normal and uniform distributions, respectively):

- We use two types of test matrix for A:
  Type 1: $A = \tilde{A}/\|\tilde{A}\|_F$, $\tilde{A} = \mathtt{randn}(100, 40)$. Typically $\kappa_2(A) \le 10$.
  Type 2: $A = \tilde{A}/\|\tilde{A}\|_F$, $\tilde{A} = U\Sigma V^T$, 40×40 $\Sigma = \mathrm{diag}(\sigma_i)$, $\sigma_i = 10^{-4(i-1)/39}$, $U \in \mathbb{R}^{100\times 40}$ and $V \in \mathbb{R}^{40\times 40}$ are the Q-factors of the QR factorizations of two random matrices $\mathtt{randn}(100, 40)$ and $\mathtt{randn}(40, 40)$, respectively. Note that $\kappa_2(A) = 10^4$.
- $b = (A + E)x$, $x = [1, \ldots, 1]^T \in \mathbb{R}^{40}$, $E = \frac{\delta_A}{\sqrt{100\times 40}}\mathtt{rand}(100, 40)$ (note that $\|E\|_F \le \delta_A$), $\delta_A = 10^{-7}, 10^{-6}, \ldots, 10^{-1}$ for type 1 matrices A, $\delta_A = 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}$ for type 2 matrices A. The DLS estimate usually has no accurate digits compared with x if δ_A is taken to be larger.
- $y = \hat{x} + \frac{1}{\sqrt{40}}\delta_{\hat{x}}\|\hat{x}\|_2\,\mathtt{rand}(40, 1)$, $\hat{x}$ is the computed solution to the DLS problem (2.1) and $\delta_{\hat{x}} = 0, 10^{-7}, 10^{-6}, \ldots, 10^{-1}$.
- For each pair of δ_A and $\delta_{\hat{x}}$ and each type of matrix, we generated 1000 sample problems.

The solution $\hat{x}$ to the DLS problem satisfies (see [5]):

$$\hat{x} = \frac{b^T b}{b^T A v_{\scriptscriptstyle D}} v_{\scriptscriptstyle D}, \tag{5.1}$$

where $v_{\scriptscriptstyle D}$ is the right singular vector corresponding to the smallest singular value of $(I - bb^{\dagger})A$. The equality (5.1) can also be obtained from [21, Sec. 9], which suggests a way to compute $\hat{x}$. In our numerical tests we used the MATLAB built-in function `svd` to find $v_{\scriptscriptstyle D}$ and then computed $\hat{x}$. To compute the asymptotic estimate $\tilde{\mu}_F(y)$, we first computed the QR factorization of B (see (4.6)) to find the Q-factor and then used (4.9).

In our numerical tests single precision was used to generate the data A, b and y, and to compute the DLS solution $\hat{x}$; and both single precision and double precision were used to compute both sides of the inequality in (2.31) (where ΔA gives the minimum). The number of failures to satisfy the inequality for each case by single (S) and double (D) precision is reported in Table 5.1 for type 1 matrices, and in Table 5.2 for type 2 matrices. When δ_A is small or $\delta_{\hat{x}}$ is large, we see that the computed version of the inequality by single precision sometimes fails. In particular for ill-conditioned type 2 matrices, when $\delta_A = 10^{-7}$ or $\delta_{\hat{x}} = 10^{-1}$, the failure percentage is very large. However the computed version of the inequality by double precision *always* holds for these test cases. This shows that these failures were due to rounding errors in the single precision computed version of the inequality, and for these test cases the extended minimal backward error is actually the true minimal backward error. The reason single precision rounding errors caused some tests to fail is almost certainly the following: in each failed case the gap between the smallest and the second smallest singular values of N_∞ was small, making the computation of the singular vector w_* (see (2.49)) inaccurate (see, e.g., [2, Thm 1.2.8] or [8, Thm 8.6.5] for perturbation results concerning the singular vectors). Indeed we noticed that for the failed cases w_* computed by single precision was very inaccurate compared with the one computed by double precision, leading to a large computational error in $\widehat{\Delta A}$, where this is needed for checking the inequality in (2.31).

In Figures 5.1 to 5.8 we give the plots corresponding to eight extreme cases in Tables 5.1 and 5.2 which exhibit $\mu_F(y)$ (as abscissa) vs $\mu_2^{\mathtt{lb}}(y)$ in (3.2), and $\mu_F(y)$ (as abscissa) vs $\tilde{\mu}_F(y)$ in (4.2) and (4.9), represented by the points $\cdot$ (blue) for $\mu_2^{\mathtt{lb}}(y)$, and $*$ (green) for $\tilde{\mu}_F(y)$. The diagonal (red) is plotted for reference. In these figures the above quantities were computed by double precision. But we can see no difference between these figures and the corresponding ones obtained by single precision.

From Figures 5.1, 5.2, 5.5 and 5.6, where each y is the computed DLS solution $\hat{x}$, we see that the minimal backward error $\mu_F(\hat{x}) \approx 10^{-7}$, which is close to the unit roundoff for single precision, so that each computed $\hat{x}$ is a backward stable solution. It is interesting to see from Figures 5.3, 5.4, 5.7 and 5.8 that $\mu_F(y)$ is about one or two orders of magnitude smaller than $\delta_{\hat{x}}$. This phenomenon also holds for other test cases.

All these figures and the figures we did not display here indicate that the lower bound $\mu_2^{\mathtt{lb}}(y)$ is a reasonable approximation to the minimal backward error $\mu_F(y)$ in the sense that these two always had the same order of magnitude, although the case for type 1 matrices is worse than the case for type 2 matrices. We also see that the asymptotic estimate $\tilde{\mu}_F(y)$ is an excellent approximation to $\mu_F(y)$, even when y is not close to the DLS solution $\hat{x}$, see Figures 5.3, 5.4, 5.7, and 5.8, where $\delta_{\hat{x}} = 10^{-1}$.

			$\delta_{\hat{x}}$							
			0	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
δ_A	10^{-7}	S	9	2	4	2	1	1	2	1
		D	0	0	0	0	0	0	0	0
	10^{-6}	S	4	2	1	2	4	1	4	1
		D	0	0	0	0	0	0	0	0
	10^{-5}	S	0	0	0	0	0	4	3	0
		D	0	0	0	0	0	0	0	0
	10^{-4}	S	0	0	0	0	0	0	2	1
		D	0	0	0	0	0	0	0	0
	10^{-3}	S	0	0	0	0	0	0	0	0
		D	0	0	0	0	0	0	0	0
	10^{-2}	S	0	0	0	0	0	0	0	0
		D	0	0	0	0	0	0	0	0
	10^{-1}	S	0	0	0	0	0	0	0	0
		D	0	0	0	0	0	0	0	0

TABLE 5.1
Number of failures to satisfy the inequality (2.31) *out of 1000 samples for type 1*

			$\delta_{\hat{x}}$							
			0	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
δ_A	10^{-7}	S	718	725	726	749	965	997	1000	997
		D	0	0	0	0	0	0	0	0
	10^{-6}	S	2	0	1	0	107	970	1000	997
		D	0	0	0	0	0	0	0	0
	10^{-5}	S	0	0	0	0	0	120	945	995
		D	0	0	0	0	0	0	0	0
	10^{-4}	S	0	0	0	0	0	0	123	888
		D	0	0	0	0	0	0	0	0

TABLE 5.2
Number of failures to satisfy the inequality (2.31) *out of 1000 samples for type 2*

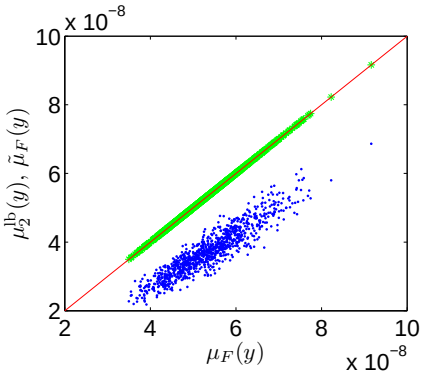

FIG. 5.1. *Type 1 A,* $\delta_A = 10^{-7}$, $\delta_{\hat{x}} = 0$

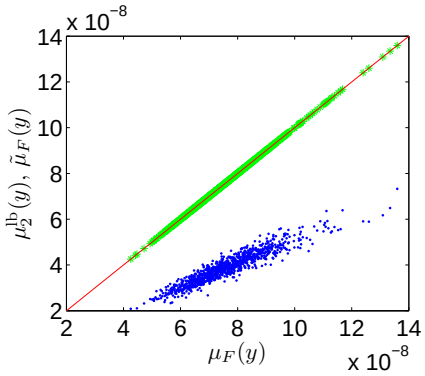

FIG. 5.2. *Type 1 A,* $\delta_A = 10^{-1}$, $\delta_{\hat{x}} = 0$

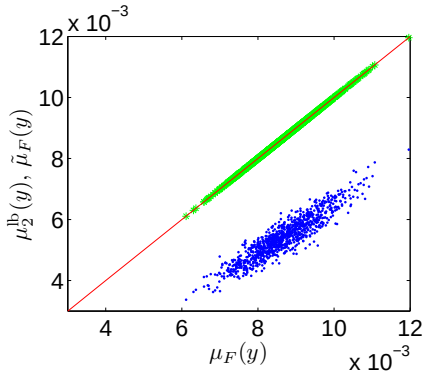

FIG. 5.3. *Type 1 A*, $\delta_A = 10^{-7}$, $\delta_{\hat{x}} = 10^{-1}$

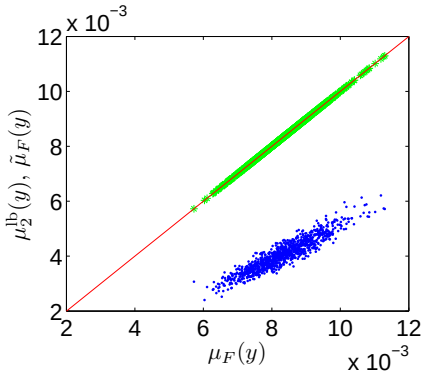

FIG. 5.4. *Type 1 A*, $\delta_A = 10^{-1}$, $\delta_{\hat{x}} = 10^{-1}$

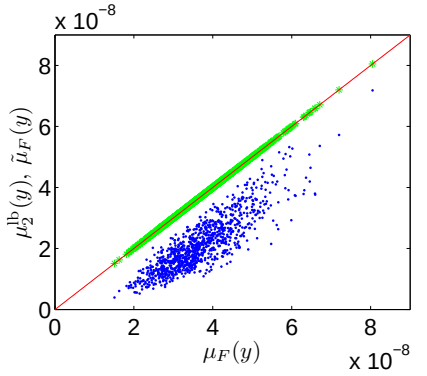

FIG. 5.5. *Type 2 A*, $\delta_A = 10^{-7}$, $\delta_{\hat{x}} = 0$

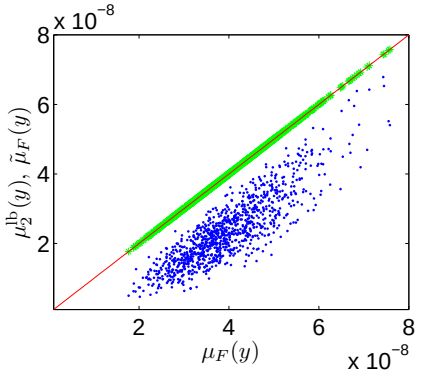

FIG. 5.6. *Type 2 A*, $\delta_A = 10^{-4}$, $\delta_{\hat{x}} = 0$

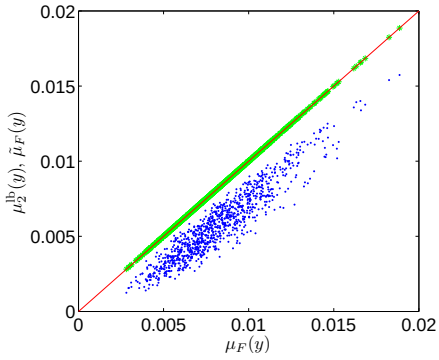

FIG. 5.7. *Type 2 A*, $\delta_A = 10^{-7}$, $\delta_{\hat{x}} = 10^{-1}$

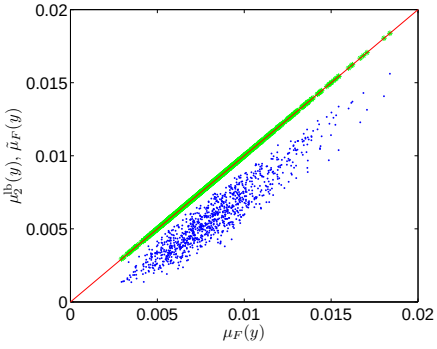

FIG. 5.8. *Type 2 A*, $\delta_A = 10^{-4}$, $\delta_{\hat{x}} = 10^{-1}$

6. Summary and future work. For a given approximate solution y to the DLS problem (2.1), we first presented formulas (2.15) in Theorem 2.2 for an extended minimal backward error $\mu_F(y, \theta)$ for the case where backward perturbations in both A and b are allowed. Then by taking $\theta \to \infty$ we obtained the corresponding formulas

(2.48) in Theorem 2.6 for an extended minimal backward error $\mu_F(y)$ for the case where only backward perturbations in A are allowed—this is the case we considered later in the paper. In theory $\mu_F(y)$ is a lower bound on the minimal backward error, but if the inequality in (2.31) is satisfied for the optimal perturbation $\widehat{\Delta A}$ given in (2.49), it is in fact the the minimal backward error. Our simulations showed that if y is a reasonable approximation to the exact DLS solution (that is, having a relative error in y of less than 10^{-1} for our test cases), then, apparently the inequality in (2.31) holds in the absence of rounding errors in checking the inequality. Thus we believe in practice that $\mu_F(y)$ can usually be used as the minimal backward error. Since the formula (2.48) for $\mu_F(y)$ involves the minimum singular value of a matrix, it is expensive to compute directly. In order to overcome this problem we derived a lower bound $\mu_2^{\mathtt{lb}}(y)$ (see (3.2)) and an asymptotic estimate $\tilde{\mu}_F(y)$ (see (4.6), (4.8) and (4.9)). These can be computed or estimated more efficiently. For our numerical test cases $\mu_2^{\mathtt{lb}}(y)$ always had the same order of magnitude as $\mu_F(y)$, and $\tilde{\mu}_F(y)$ was an excellent approximation to $\mu_F(y)$. Since the computation of $\mu_2^{\mathtt{lb}}(y)$ is so inexpensive, it would seem to give a simple and effective indicator.

Several problems need to be investigated in the future. To check if the extended minimal backward error is the actual minimal backward error we need an efficient and reliable way to test the inequality in (2.31) (or the inequality in (2.9) when perturbations in both A and b are allowed). The relationships between $\mu_F(y)$ and $\tilde{\mu}_F(y)$ needs to be studied further. We would also like to incorporate the results obtained in this paper to design effective stopping criteria for iterative algorithms for solving the DLS problem, and extend the results here to total least squares problems (see [7] and [33]) and scaled total least squares problems (see [23] and [21]).

Acknowledgments. The first author is indebted to the second author and to Michael Saunders for their hospitality when he was on sabbatical at Stanford University where part of this research was done. He is also grateful to Zheng Su for helpful discussions. We also thank Pete Stewart and Nick Trefethen for improving our understanding of eigensystems of analytic Hermitian matrices. Two referees made many helpful suggestions, some of which were very demanding but led to necessary and substantial expansions of this paper. One referee also suggested the form of Theorem 2.8 in order to give a clearer exposition of our thinking.

REFERENCES

[1] M. Arioli, I. Duff and D. Ruiz, *Stopping criteria for iterative solvers*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 138144.
[2] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, PA, 1996.
[3] X.-W. Chang, C. C. Paige and D. Titley-Peloquin, *Characterizing matrices that are consistent with given solutions*. Submitted for publication.
[4] A.J. Cox and N.J. Higham, *Backward error bounds for constrained least squares problems*, BIT, 39 (1999), pp. 210–227.
[5] R.D. DeGroat and E.M. Dowling, *The data least squares problem and channel equalization*, IEEE Trans. Signal Processing, 42 (1993), pp. 407–411.
[6] G.H. Golub and G. Meurant, *Matrices, moment and quadrature*, in Proceedings of the 15th Dundee Conference, June-July 1993, D. F. Griffiths and G. A. Watson (eds.), Longman Scientific & Technical, 1994.
[7] G.H. Golub and C.F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17:883–893, 1980.
[8] G.H. Golub and C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore MD, third ed., 1996.

[9] J.F. Grcar, *Optimal sensitivity analysis of linear least squares*, Technical Report LBNL-52434. Lawrence Berkeley National Laboratory, 2003.
[10] J.F. Grcar, *The optimum inverse problem of numerical error analysis*, Householder Symposium XVI, Seven Springs, USA, May 23-27, 2005.
[11] J.F. Grcar, M.A. Saunders, and Z. Su, *Estimates of optimal backward perturbations for linear least squares problems*, manuscript, 2004.
[12] M. Gu, *Backward perturbation bounds for linear least squares problems*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 363–372.
[13] N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd Edn., SIAM, Philadelphia, PA, 2002.
[14] D.J. Higham and N. J. Higham, *Backward error and condition of structured linear systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 162–175.
[15] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[16] T. Kato, *Perturbation Theory for Linear Operators*, 2nd Edn., Springer-Verlag, New York, 1980.
[17] R. Karlson and B. Waldén, *Estimation backward perturbation bounds for the linear least squares problem*, BIT, 37 (1997), 862-869.
[18] A.N. Malyshev, *Optimal backward perturbation bounds for the LSS problems*, BIT, 41 (2001), 430–432.
[19] A.N. Malyshev and M. Sadkane, *Computation of optimal backward perturbation bounds for large sparse linear least squares problems*, BIT, 41 (2002), pp. 739–747.
[20] C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8:43–71, 1982.
[21] C.C. Paige and Z. Strakoš, *Scaled total least squares fundamentals*, Numerische Mathematik, 91:117–146 (2002).
[22] C.C. Paige and Z. Strakoš, *Unifying least squares, total least squares and data least squares*, in "Total Least Squares and Errors-in-Variables Modeling", S. van Huffel and P. Lemmerling, editors, Kluwer Academic Publishers, Dordrecht, 2002, pp. 25–34.
[23] B.D. Rao, *Unified treatment of LS, TLS and truncated SVD methods using a weighted TLS framework*. In: S. Van Huffel (editor), *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modelling*, pp. 11–20, SIAM Publications, Philadelphia PA, 1997.
[24] F. Rellich, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, 1969.
[25] J.L. Rigal and J. Gaches, *On the compatibility of a given solution with the data of a linear system. JACM*, 14:543–548 (1967).
[26] G.W. Stewart, *On the perturbation of pseudo-inverses, projections, and linear least squares problems. SIAM Rev.*, 19:634–662 (1977).
[27] Z. Su, *Computational Methods for Least Squares Problems and Clinical Trials*. Ph.D Thesis, Scientific Computing & Computational Mathematics, Stanford University, 2005.
[28] J.-G. Sun, *Optimal backward perturbation bounds for the linear least squares problem with multiple right-hand sides*, IMA J. Numer. Anal., 16 (1996), pp. 1–11.
[29] J.-G. Sun, *On optimal backward perturbation bounds for the linear least-squares problem*, BIT, 37 (1997), pp. 179–188.
[30] J.-G. Sun, Bounds for the structured backward errors of Vandermonde systems, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 45–59.
[31] J.-G. Sun, *A note on backward errors for structured linear systems*, Numerical Linear Algebra with Applications, 12(7), 2005, pp 585-603.
[32] J.-G. Sun and Z. Sun, *Optimal backward perturbation bounds for underdetermined systems*. SIAM J. Matrix Anal. Appl., 18 (1997), pp. 393–402.
[33] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM Publications, Philadelphia PA, 1991.
[34] J.M. Varah, *Backward error estimates for Toeplitz systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 408–417.
[35] B. Waldén, R. Karlson and J.-G. Sun, *Optimal backward perturbation bounds for the linear least squares problem*, Numerical Linear Algebra with Applications, 2:271–286 (1995).
[36] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK (1965).
[37] H. Xiang and Y. Wei *On normwise structured backward errors for saddle point systems*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 838-849.