Improved Error Bounds for Underdetermined System Solvers

James W. Demmel [*]   Nicholas J. Higham [†]

July 2, 1991

Abstract

The minimal 2-norm solution to an underdetermined system $Ax = b$ of full rank can be computed using a QR factorization of A^T in two different ways. One requires storage and re-use of the orthogonal matrix Q while the method of semi-normal equations does not. Existing error analyses show that both methods produce computed solutions whose normwise relative error is bounded to first order by $c\kappa_2(A)u$, where c is a constant depending on the dimensions of A, $\kappa_2(A) = \|A^+\|_2\|A\|_2$ is the 2-norm condition number, and u is the unit roundoff. We show that these error bounds can be strengthened by replacing $\kappa_2(A)$ by the potentially much smaller quantity $\mathrm{cond}_2(A) = \|\,|A^+|\cdot|A|\,\|_2$, which is invariant under row scaling of A. We also show that $\mathrm{cond}_2(A)$ reflects the sensitivity of the minimum norm solution x to row-wise relative perturbations in the data A and b. For square linear systems $Ax = b$ row equilibration is shown to endow

[*]Computer Science Division and Mathematics Department, University of California, Berkeley, CA 94720, U.S.A. (`na.demmel@na-net.ornl.gov`). This author acknowledges the financial support of the National Science Foundation via grants DCR-8552474 and ASC-8715728. He is also a Presidential Young Investigator.

[†]Department of Mathematics, University of Manchester, Manchester, M13 9PL, UK. (`na.nhigham@na-net.ornl.gov`).

solution methods based on LU or QR factorization of A with relative error bounds proportional to $\mathrm{cond}_\infty(A)$, just as when a QR factorization of A^T is used. The advantages of using fixed precision iterative refinement in this context instead of row equilibration are explained.

Key words: underdetermined system, semi-normal equations, QR factorization, rounding error analysis, backward error, componentwise error bounds, iterative refinement, row scaling.

AMS(MOS) subject classifications. primary 65F05, 65F25, 65G05.

Appeared in SIAM J. Mat. Anal. Appl. v. 14, n. 1, 1993. Also appeared as LAPACK Working Note 23.

1 Introduction

Consider the underdetermined system $Ax = b$, where $A \in \mathbb{R}^{m\times n}$ with $m \leq n$. The system can be analysed using a QR factorization

$$A^T = Q \begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{1.1}$$

where $Q \in \mathbb{R}^{n\times n}$ is orthogonal and $R \in \mathbb{R}^{m\times m}$ is upper triangular. We have

$$b = Ax = [\, R^T \quad 0 \,]\, Q^T x = R^T y_1, \tag{1.2}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = Q^T x.$$

If A has full rank then $y_1 = R^{-T} b$ is uniquely determined and all solutions of $Ax = b$ are given by

$$x = Q \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \qquad y_2 \in \mathbb{R}^{n-m} \text{ arbitrary.}$$

The unique solution x_{LS} that minimizes $\|x\|_2$ is obtained by setting $y_2 = 0$. We have

$$\begin{aligned} x_{LS} &= Q \begin{bmatrix} R^{-T} b \\ 0 \end{bmatrix} && (1.3) \\ &= Q \begin{bmatrix} R \\ 0 \end{bmatrix} R^{-1} R^{-T} b = Q \begin{bmatrix} R \\ 0 \end{bmatrix} (R^T R)^{-1} b && (1.4) \\ &= A^T (A A^T)^{-1} b \\ &= A^+ b, \end{aligned}$$

where $A^+ = A^T(AA^T)^{-1}$ is the pseudo-inverse of A.

Equation (1.3) defines one way to compute x_{LS}. This is the method described in [13, Ch. 13], and we will refer to it as the "Q method". When A is large and sparse it is desirable to avoid storing and accessing Q, which can be expensive. An alternative method with this property was suggested by Gill and Murray [6] and Saunders [16]. This method again uses the QR factorization (1.1) but computes x_{LS} as

$$x_{LS} = A^T y$$

where

$$R^T R y = b \qquad (1.5)$$

(cf. (1.4)). These latter equations are called the semi-normal equations (SNE), since they are equivalent to the "normal equations" $AA^T y = b$. As the "semi" denotes, however, this method does not explicitly form AA^T, which would be undesirable from the standpoint of numerical stability. We stress that equations (1.5) are different from the equations $R^T R x = A^T b$ for an overdetermined least squares problem, where $A = Q\,[\,R^T \quad 0\,]^T \in \mathbb{R}^{m \times n}$ with $m \geq n$, yet these are also referred to as semi-normal equations [4]. In this paper we are solely concerned with underdetermined systems so no confusion should arise.

Other methods for obtaining minimal 2-norm solutions of underdetermined systems are surveyed in [5].

Existing perturbation theory for the minimum norm solution problem, and error analysis for the above QR factorization-based methods, can be summarised as follows.

(1) Golub and Van Loan [7, Th. 5.7.1] prove the following perturbation result. (Similar results are proved in [13, Th. 9.18] and [20, Th. 5.1]. Here, $\sigma_i(A)$ denotes the ith largest singular value of $A \in \mathbb{R}^{m\times n}$ and, if $\mathrm{rank}(A) = m$, $\kappa_2(A) = \|A^+\|_2\|A\|_2 = \sigma_1(A)/\sigma_m(A)$.

Theorem 1.1 *Let $A \in \mathbb{R}^{m\times n}$ and $0 \neq b \in \mathbb{R}^m$. Suppose $\mathrm{rank}(A) = m \leq n$ and that $\Delta A \in \mathbb{R}^{m\times n}$ and $\Delta b \in \mathbb{R}^m$ satisfy*

$$\epsilon = \max\{\|\Delta A\|_2/\|A\|_2, \|\Delta b\|_2/\|b\|_2\} < \sigma_m(A).$$

If x and $\widehat{x}$ are the minimum norm solutions to $Ax = b$ and $(A + \Delta A)\widehat{x} = b + \Delta b$ respectively, then

$$\frac{\|\widehat{x} - x\|_2}{\|x\|_2} \leq \min\{3, n - m + 2\}\kappa_2(A)\epsilon + O(\epsilon^2). \tag{1.6}$$

This result shows that small relative changes in the data A and b produce relative changes in the minimum norm solution x that are at most $\kappa_2(A)$ times as large. Unlike for the overdetermined least squares problem there is no term in $\kappa_2(A)^2$.

(2) Arioli and Laratta [2, Th. 4] show that the computed solution $\widehat{x}$ from the Q method satisfies

$$\frac{\|\widehat{x} - x\|_2}{\|x\|_2} \leq c_1 u\kappa_2(A) + O(u^2), \tag{1.7}$$

where c_i denotes a modest constant depending on m and n, and u is the unit roundoff. (Arioli and Laratta actually analyse a slightly more general problem in which $\|x - w\|_2$ is minimized for a given vector w; we have taken $w = 0$).

(3) Paige [15] shows that the computed solution $\widehat{x}$ from the method of semi-normal equations satisfies

$$\frac{\|\widehat{x}-x\|_2}{\|x\|_2} \le \kappa_2(A)c_1u + \frac{\kappa_2(A)c_2u(1+\kappa_2(A)c_3u)}{1-\kappa_2(A)c_4u}. \tag{1.8}$$

The bounds (1.7) and (1.8) are of the same form as (1.6). One implication of these existing results is that both the Q method and the SNE method are stable in the sense that the relative errors in the computed solutions reflect the sensitivity of the minimum norm problem to general perturbations in the data.

The purpose of this paper is to show that the results in (2) and (3) can be strengthened significantly by employing componentwise analysis. First, in section 2, we prove a version of Theorem 1.1 for componentwise perturbations; thus we measure ΔA and Δb by the smallest ϵ such that

$$|\Delta A| \le \epsilon E, \quad |\Delta b| \le \epsilon f, \tag{1.9}$$

where $E \ge 0$ and $f \ge 0$ contain arbitrary tolerances and inequalities hold componentwise. We obtain an analogue of (1.6) with $\kappa_2(A)$ replaced by a potentially much smaller quantity that depends on A, x, E and f.

In section 3 we show that the term $\kappa_2(A)$ in (1.7) and (1.8) can be replaced by

$$\mathrm{cond}_2(A) \equiv \|\, |A^+| \cdot |A| \,\|_2,$$

which is a generalization of the condition number $\|\, |A^{-1}| \cdot |A| \,\|_2$ for square matrices introduced by Bauer [3] and Skeel [17]. This is important because $\mathrm{cond}_2(A)$ can be arbitrarily smaller than $\kappa_2(A)$, since $\mathrm{cond}_2(A)$ is invariant under row scalings $A \to DA$ (D diagonal and nonsingular) whereas $\kappa_2(A)$ is not. And $\mathrm{cond}_2(A)$ cannot be much bigger than $\kappa_2(A)$ since

$$\mathrm{cond}_2(A) \le \|\, |A^+| \,\|_2 \|\, |A| \,\|_2 \le n\|A^+\|_2\|A\|_2 = n\kappa_2(A). \tag{1.10}$$

In sections 4 and 5 we investigate stability issues, and we encounter several different types of stability. To put these different types into perspective

Table 1.1: Stability classification scheme.

	Backward stability	Forward stability
Normwise	N	N
Row-wise	R	R
Componentwise	C	C

we present in Table 1.1 a scheme that classifies six different kinds of stability. (We appreciate that it can be counter-productive to over-formalize stability, but we believe that this scheme helps to clarify the overall picture.)

To explain the terminology we define for $A \in \mathbb{R}^{m\times n}$, with $m \le n$, the backward error

$$\begin{aligned} \omega_{E,f}(y) \quad \equiv \quad & \min\{\epsilon : \exists\, \Delta A \in \mathbb{R}^{m\times n}, \Delta b \in \mathbb{R}^m \text{ s.t. } y \text{ is the minimum norm} \\ & \quad \text{solution to } (A+\Delta A)y = b + \Delta b, \text{ and } |\Delta A| \le \epsilon E,\ |\Delta b| \le \epsilon f\}, \end{aligned}$$

where $E \ge 0$ and $f \ge 0$ are given. Note that if we were to remove the minimum norm requirement on y in the definition of $\omega_{E,f}$ then the backward error would be given by

$$\max_i \frac{|b - Ay|_i}{(E|y| + f)_i}, \tag{1.11}$$

as shown in [14]. The three measures of backward stability in Table 1.1 correspond to the following choices of E and f, where $e_n = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$:

$$\begin{aligned} \text{normwise } (\omega^N): &\qquad E_N = \|A\|_2 e_m e_n^T, \quad f_N = \|b\|_2 e_m, \\ \text{row-wise } (\omega^R): &\qquad E_R = |A| e_n e_n^T, \quad f_R = |b|, \\ \text{componentwise } (\omega^C): &\qquad E_C = |A|, \quad f_C = |b|. \end{aligned} \tag{1.12}$$

A small value for $\omega^R(y)$ means that y is the minimum norm solution to a perturbed system where the perturbation to the ith row of A is small compared with the norm of the ith row (similarly for b). We say, for example, that a numerical method for solving $Ax = b$ is in backward stability category R (or is row-wise backward stable) if it produces a computed solution $\widehat{y}$ such that $\omega^R(\widehat{y})$ is of order the unit roundoff

For each type of backward error there is a perturbation result that bounds $\|x - y\|_2/\|x\|_2$ by a multiple of $\omega_{E,f}(y)$, and the multiplier defines a condition number. As explained in section 2, for underdetermined systems the conditions numbers are $\kappa_2(A)$ for ω^N, $\mathrm{cond}_2(A)$ for ω^R, and a quantity $\mathrm{cond}_2(A, x)$ that depends on both A and x for ω^C. Continuing the "R-stability" example above, we say that a method is in forward stability category R if it has a forward error bound of order $\mathrm{cond}_2(A)$ times the unit roundoff. An algorithm that has backward stability X (where $X = N$, R, or C) automatically has forward stability X; one of the reasons these definitions are useful is that an algorithm can have forward stability X without having backward stability X.

In this terminology, the gist of section 3 is that the Q method and the SNE method have forward stability R, whereas previous results guaranteed only forward stability N.

In section 4 we explain why the Q method is (nearly) row-wise backward stable but the SNE method is not backward stable at all. We give some numerical results to provide insight into the error bounds and to illustrate the performance of fixed precision iterative refinement with the SNE method.

In section 5 we consider the implications of the results of section 3 for square linear systems. We show that row equilibration of the system $Ax = b$ allows methods based on LU and QR factorization of A to produce computed solutions whose relative errors are bounded in the same way as when a QR factorization of A^T is employed—namely by a multiple of $\mathrm{cond}(A)u$ (corresponding to row-wise forward stability). We explain why fixed precision iterative refinement leads to an even more satisfactory computed solution than row equilibration and we provide two numerical examples for illustration.

2 Componentwise Perturbation Result

In this section we prove the following componentwise perturbation result for the minimum norm problem, and use it to determine the condition numbers for the perturbation measures in (1.12).

Theorem 2.1 *Let $A \in \mathbb{R}^{m\times n}$ and $0 \neq b \in \mathbb{R}^m$. Suppose* $\operatorname{rank}(A) = m \leq n$, *and that*

$$|\Delta A| \leq \epsilon E, \quad |\Delta b| \leq \epsilon f,$$

where $E \geq 0$, $f \geq 0$, and $\epsilon\|E\|_2 < \sigma_m(A)$. If x and $\widehat{x}$ are the minimum norm solutions to $Ax = b$ and $(A + \Delta A)\widehat{x} = b + \Delta b$ respectively, then

$$\frac{\|\widehat{x} - x\|_2}{\|x\|_2} \leq \Big(\|\, |I - A^+A| \cdot E^T \cdot |A^{+^T} x|\, \|_2 + \|\, |A^+| \cdot (f + E|x|)\, \|_2\Big) \frac{\epsilon}{\|x\|_2} + O(\epsilon^2). \tag{2.1}$$

Proof. $A + \Delta A$ has full rank so we can manipulate the equation

$$\widehat{x} = (A + \Delta A)^T((A + \Delta A)(A + \Delta A)^T)^{-1}(b + \Delta b)$$

to obtain

$$\begin{aligned} \widehat{x} - x &= (I - A^+A)\Delta A^T(AA^T)^{-1}b + A^+(\Delta b - \Delta Ax) + O(\epsilon^2) \\ &= (I - A^+A)\Delta A^T A^{+^T} x + A^+(\Delta b - \Delta Ax) + O(\epsilon^2). \end{aligned} \tag{2.2}$$

Taking norms and then using absolute value inequalities, together with the monotonicity property $|x| \leq y \Rightarrow \|x\|_2 \leq \|y\|_2$, we have

$$\begin{aligned} \|\widehat{x} - x\|_2 &\leq \|(I - A^+A)\Delta A^T A^{+^T} x\|_2 + \|A^+(\Delta b - \Delta Ax)\|_2 + O(\epsilon^2) \\ &\leq (\|\, |I - A^+A| \cdot E^T \cdot |A^{+^T} x|\, \|_2 + \|\, |A^+| \cdot (f + E|x|)\, \|_2)\epsilon + O(\epsilon^2 \end{aligned}$$

as required. ∎

We note that for given A, b, E and f there exist ΔA and Δb for which the bound in (2.1) is attained to within a constant factor depending on n. This is a consequence of the fact that the two vectors on the right-hand side of (2.2) are orthogonal. Also, it is clear from the proof that (2.1) is valid with the 2-norm replaced by the ∞-norm.

By substituting the E and f from (1.12) into Theorem 2.1 we can deduce the condition numbers corresponding to our three different ways of measuring the perturbations ΔA and Δb. For the componentwise measure the condition number is clearly

$$\mathrm{cond}_2(A,x) = (\|\,|I - A^+A|\cdot|A^T|\cdot|A^{+^T}x|\,\|_2 + \|\,|A^+|\cdot(|b| + |A||x|)\,\|_2)/\|x\|_2. \tag{2.3}$$

Replacing b by its upper bound $|A||x|$ simplifies this expression while increasing it by no more than a factor of 2.

For the row-wise measure the bracketed term in the bound in (2.1) is within a factor depending on n of $\mathrm{cond}_2(A)$, hence we can take $\mathrm{cond}_2(A)$ as the condition number. In showing this one needs to use the equality $\|I - A^+A\|_2 = \min\{1, n-m\}$ (which can be derived by consideration of the QR factorization (1.1), for example), and the observation that if $B \in \mathbb{R}^{m\times n}$ and $B \geq 0$ then

$$\frac{1}{\sqrt{m}}\|B\|_2 \leq \|B\|_\infty = \|Be\|_\infty \leq \|Be\|_2 \leq \sqrt{n}\|B\|_2.$$

Note that when $|x| = e$, $\mathrm{cond}_2(A)$ differs from $\mathrm{cond}_2(A,x)$ by no more than a factor of about $\sqrt{n}$. Finally, for the normwise measure the condition number is $\kappa_2(A)$ (as implied by Theorem 1.1). Table 2.1 summarises these results.

In the error analysis of the next section we need to use Theorem 2.1 with $E = |A|H$, where H is a given matrix. In this case, taking also $f = |b|$, it is convenient to put (2.1) in the form

$$\frac{\|\widehat{x} - x\|_2}{\|x\|_2} \leq \min\{3, n-m+2\}\max\{\|H\|_2, 1\}\mathrm{cond}_2(A)\epsilon + 0(\epsilon^2). \tag{2.4}$$

Table 2.1: Condition Numbers.

Measure	Condition Number
Normwise	$\kappa_2(A)$
Row-wise	$\mathrm{cond}_2(A)$
Componentwise	$\mathrm{cond}_2(A, x)$

If $\|H\|_2 = 1$ this is precisely (1.6) with $\kappa_2(A)$ replaced by $\mathrm{cond}_2(A)$, this difference reflecting the stronger assumption made about the perturbations for (2.4).

3 Error Analysis

In this section we carry out an error analysis of the Q method and the SNE method. We assume that the floating point arithmetic obeys the model

$$\begin{aligned} fl(x \,\mathrm{op}\, y) &= (x \,\mathrm{op}\, y)(1+\delta), \qquad |\delta| \le u, \quad \mathrm{op} = *, /, \\ fl(x \pm y) &= x(1+\alpha) \pm y(1+\beta), \qquad |\alpha|, |\beta| \le u, \\ fl(\sqrt{x}) &= \sqrt{x}(1+\delta), \qquad |\delta| \le u. \end{aligned}$$

We consider first the Q method, and we assume that the QR factorization (1.1) is computed by Householder transformations or Givens transformations. In [12, Cor. A.8] it is shown that if $\widehat{R}$ is the computed upper triangular factor there exists an orthogonal matrix $\widetilde{Q}$ such that

$$A^T + \Delta A^T = \widetilde{Q} \begin{bmatrix} \widehat{R} \\ 0 \end{bmatrix}, \tag{3.1}$$

where

$$|\Delta A^T| \le G_{m,n} u |A^T| \tag{3.2}$$

and $\|G_{m,n}\|_2 \le \mu_{m,n}$. Here and below we use $\mu_{m,n}$ generically to denote a modest constant depending on m and n; we are not concerned with the precise values of the constants so will freely write, for example, $\mu_{m,n}+\mu'_{m,n} = \mu''_{m,n}$.

The Q method solves the triangular system $R^T y_1 = b$ and forms $x = Q\,[\,y_1^T,\ 0\,]^T$. Standard analysis shows that the computed $\widehat{y}_1$ satisfies

$$(\widehat{R} + \Delta\widehat{R})^T \widehat{y}_1 = b, \quad |\Delta\widehat{R}| \le \mu_m u |\widehat{R}|. \tag{3.3}$$

¿From [12, Lemma A.7] the computed solution $\widehat{x}$ satisfies

$$\widehat{x} = \widetilde{Q} \begin{bmatrix} \widehat{y}_1 \\ 0 \end{bmatrix} + g, \tag{3.4}$$

where

$$|g| \le G'_{m,n} \begin{bmatrix} |\widehat{y}_1| \\ 0 \end{bmatrix} u, \quad \|G'_{m,n}\|_2 \le \mu'_{m,n}. \tag{3.5}$$

(We emphasise the important point that the same orthogonal matrix $\widetilde{Q}$ appears in (3.1) and (3.4).)

Ideally, we would like to use the basic error equations (3.1)–(3.5) to show that $\widehat{x}$ is the exact minimum norm solution to a perturbed problem where the perturbations are bounded according to $|\Delta A| \le \epsilon|A|$ and $|\Delta b| \le \epsilon|b|$. The forward error could then be bounded by invoking (2.1). Unfortunately, this componentwise backward stability result does not hold. We can, nevertheless, obtain a forward error bound of the form (2.4) by using a mixed forward and backward error argument.

¿From (3.3), (3.4) and (3.1) we have

$$\begin{aligned} b &= [\,(\widehat{R} + \Delta\widehat{R})^T \quad 0\,]\,\widetilde{Q}^T \cdot \widetilde{Q} \begin{bmatrix} \widehat{y}_1 \\ 0 \end{bmatrix} \\ &= (A + F)\overline{x}, \end{aligned}$$

where

$$\begin{aligned} F &= \Delta A + [\,\Delta\widehat{R}^T \quad 0\,]\,\widetilde{Q}^T, \\ \overline{x} &= \widetilde{Q} \begin{bmatrix} \widehat{y}_1 \\ 0 \end{bmatrix}. \end{aligned} \tag{3.6}$$

Since $(A+F)^T$ has the QR factorization $(A+F)^T = \widetilde{Q}\,[\,(\widehat{R}+\Delta\widehat{R})^T \quad 0\,]^T$ it follows from (3.3) and (3.6) that $\overline{x}$ is the minimum norm solution to $(A+F)\overline{x} = b$ as long as $\|F\|_2 < \sigma_m(A)$ (so that $A+F$ has full rank). ¿From (3.1)–(3.3) we have

$$\begin{aligned} |F| &\le u|A|G_{m,n}^T + \mu_m u|A|(I+uG_{m,n}^T)|\widetilde{Q}||\widetilde{Q}^T| \\ &\equiv u|A|H_{m,n}. \end{aligned}$$

Hence we can invoke (2.4) to obtain

$$\frac{\|\overline{x}-x\|_2}{\|x\|_2} \le \mu_{m,n}\mathrm{cond}_2(A)u + O(u^2). \tag{3.7}$$

Now from (3.4), (3.5) and (3.6) we have

$$\begin{aligned} \|\widehat{x}-\overline{x}\|_2 &= \|g\|_2 \\ &\le \mu'_{m,n}\|\widehat{y}_1\|_2 u = \mu'_{m,n}\|\widehat{x}\|_2 u + O(u^2) \\ &= \mu'_{m,n}\|x\|_2 u + O(u^2). \end{aligned} \tag{3.8}$$

Combining (3.7) and (3.8) we conclude that

$$\frac{\|\widehat{x}-x\|_2}{\|x\|_2} \le \mu''_{m,n}\mathrm{cond}_2(A)u + O(u^2). \tag{3.9}$$

Now we analyse the SNE method. As for the Q method, (3.1) and (3.2) hold for the computed triangular factor $\widehat{R}$. The computed solution $\widehat{y}$ to (1.5) satisfies

$$(\widehat{R}+\Delta\widehat{R}_1)^T(\widehat{R}+\Delta\widehat{R}_2)\widehat{y} = b, \qquad |\widehat{R}_i| \le \mu_m u|\Delta\widehat{R}_i|, \tag{3.10}$$

and the computed solution $\widehat{x}$ satisfies

$$\widehat{x} = A^T\widehat{y} + g, \qquad |g| \le \mu_m u|A^T||\widehat{y}|. \tag{3.11}$$

Taking a similar approach to the analysis for the Q method we write

$$\widehat{x} = \overline{x} + \Delta\overline{x}, \tag{3.12}$$

where

$$\begin{aligned}
\overline{x} &= (A+\Delta A)^T \overline{y}, \\
\widehat{R}^T \widehat{R}\, \overline{y} &= b, && (3.13) \\
\Delta\overline{x} &= A^T(\widehat{y}-\overline{y}) - \Delta A^T \overline{y} + g. && (3.14)
\end{aligned}$$

Note that $\overline{x}$ is the exact minimum norm solution to $(A+\Delta A)x = b$ and so once again (3.7) holds.

For later use we note that from (3.1),

$$A + \Delta A = \widehat{R}^T \widetilde{Q}_1^T, \tag{3.15}$$

where $\widetilde{Q}_1$ comprises the first m columns of $\widetilde{Q}$, and hence, using (3.13),

$$\widetilde{Q}_1^T \overline{x} = \widehat{R}\, \overline{y}. \tag{3.16}$$

It remains to bound $\Delta\overline{x}$. Straightforward manipulation of (3.10) and (3.13) yields

$$\begin{aligned}
\overline{y} - \widehat{y} &= \widehat{R}^{-1}\widehat{R}^{-T}\Delta\widehat{R}_1^T \widehat{R}\,\overline{y} + \widehat{R}^{-1}\Delta\widehat{R}_2\,\overline{y} + O(u^2) \\
&= \widehat{R}^{-1}(\widehat{R}^{-T}\Delta\widehat{R}_1^T \widetilde{Q}_1^T \overline{x} + \Delta\widehat{R}_2\,\overline{y} + O(u^2),
\end{aligned}$$

where we have used (3.16). Pre-multiplying by A^T and using (3.15) gives

$$A^T(\overline{y}-\widehat{y}) = \widetilde{Q}_1(\widehat{R}^{-T}\Delta\widehat{R}_1^T \widetilde{Q}_1^T \overline{x} + \Delta\widehat{R}_2\,\overline{y}) + O(u^2),$$

which leads to

$$\|A^T(\overline{y}-\widehat{y})\|_2 \le \mu_m u(\|\,|\widehat{R}^{-T}|\cdot|\widehat{R}^T|\,\|_2 \|\overline{x}\|_2 + \|\,|\widehat{R}|\cdot|\overline{y}|\,\|_2) + O(u^2). \tag{3.17}$$

To bound $\|\,|\widehat{R}^{-T}|\cdot|\widehat{R}^T|\,\|_2$ note that for the exact QR factorization we have

$$\|\,|R^{-T}|\cdot|R^T|\,\|_2 = \|\,|Q_1^T A^+|\cdot|AQ_1|\,\|_2 \le m\,\mathrm{cond}_2(A).$$

Hence

$$\|\,|\widehat{R}^{-T}|\cdot|\widehat{R}^T|\,\|_2 \le m\,\mathrm{cond}_2(A+\Delta A) = m\,\mathrm{cond}_2(A) + O(u). \tag{3.18}$$

To bound $\| \, |\widehat{R}| \cdot |\overline{y}| \, \|_2$ in (3.17) we note first that for the exact R and y

$$\| \, |R| \cdot |y| \, \|_2 = \| \, |Q_1^T A^T| \cdot |y| \, \|_2 \le \sqrt{m} \| \, |A^T| \cdot |y| \, \|_2.$$

Now, since $x = A^T y$ we have $Ax = (AA^T)y$, or $y = A^{+T} x$. Hence

$$\| \, |A^T| \cdot |y| \, \|_2 \le \| \, |A^T| \cdot |A^{+T}| \cdot |x| \, \|_2 \le \mathrm{cond}_2(A) \|x\|_2. \tag{3.19}$$

It follows that for the computed $\widehat{R}$ and $\overline{y}$

$$\| \, |\widehat{R}| \cdot |\overline{y}| \, \|_2 \le \sqrt{m} \, \mathrm{cond}_2(A) \|x\|_2 + O(u). \tag{3.20}$$

Combining (3.14), (3.17), (3.18), (3.20), (3.11) and (3.19) we have

$$\|\Delta \overline{x}\|_2 \le \mu_{m,n} \mathrm{cond}_2(A) u \|x\|_2 + O(u^2).$$

Together with (3.7) and (3.12) this yields

$$\frac{\|\widehat{x} - x\|_2}{\|x\|_2} \le \mu'_{m,n} \mathrm{cond}_2(A) u + O(u^2).$$

4 Discussion and Numerical Results

The analysis in the previous section shows that for both the Q method and the SNE method the forward error is bounded by a multiple of $\mathrm{cond}_2(A)u$, so both methods are forward stable in the row-wise sense. Before giving some numerical examples we briefly consider what can be said about backward stability.

For the Q method the analysis of section 3 proves the following result about the computed solution $\widehat{x}$. There exists a vector $\overline{x}$ and a matrix F such that $\overline{x}$ is the minimum norm solution to $(A + F)\overline{x} = b$ where

$$|F| \le u|A| H_{m,n} \le \mu_{m,n} u |A| e e^T \quad \Rightarrow \quad \|F\|_2 \le \mu'_{m,n} u \|A\|_2$$

and

$$\|\widehat{x} - \overline{x}\|_2 \le \mu''_{m,n} \|\widehat{x}\|_2 u + O(u^2).$$

(This result, without the componentwise bound on F, is also proved in [13, Th. 16.18].) Thus $\widehat{x}$ is relatively close to a vector that satisfies the criterion for row-wise backward stability, and so the Q method is "almost" row-wise backward stable. Note also that, from the above, $\widehat{x}$ has a relatively small residual:

$$\|b - A\widehat{x}\|_2 \le \mu'''_{m,n}\|A\|_2\|\widehat{x}\|_2 u + O(u^2). \tag{4.1}$$

Interestingly, (4.1) implies that $\widehat{x}$ itself solves a slightly perturbed system, but it is not in general the minimum norm solution.

For the SNE method it is not even possible to derive a residual bound of the form (4.1). The method of solution guarantees only that the semi-normal equations themselves have a small residual. Thus, as in the context of overdetermined least squares problems [4] the SNE method is not backward stable.

A possible way to improve the backward stability of the SNE method is to use iterative refinement in fixed precision, as advocated in the overdetermined case in [4]. Some justification for this approach can be given using the analysis for an arbitrary linear equations solver in [12].

We have run some numerical experiments in MATLAB, which has a unit roundoff $u \approx 2.2 \times 10^{-16}$. In our experiments we rounded the result of every arithmetic operation to 23 significant bits, thus simulating single precision arithmetic with $u_{\mathrm{SP}} \approx 1.2 \times 10^{-7}$. The double precision solution was regarded as the exact solution when computing forward errors.

We report results for several 10×16 matrices A, with the right-hand sides b chosen randomly with elements from the normal $(0,1)$ distribution. We report for each approximate solution $\widehat{y}$ the normwise relative error

$$\gamma_2(\widehat{y}) = \frac{\|\widehat{y} - x\|_2}{\|x\|_2},$$

and the three relative residuals

$$\rho^X(\widehat{y}) = \max_i \frac{|b - A\widehat{y}|_i}{(E_X|\widehat{y}| + f_X)_i}, \qquad X = N, R, C,$$

where E_X and f_X are defined in (1.12). Iterative refinement in fixed precision was used with the SNE method until either $\rho^N(\widehat{y}) \leq u_{\rm SP}$ or five iterations were done. Note that if we were to use the ∞-norm in defining E_N and f_N in (1.12) then $\rho^N(\widehat{y}) \leq \rho^R(\widehat{y})$ would be guaranteed; for the 2-norm, $\rho^N(\widehat{y}) > \rho^R(\widehat{y})$ is possible. We also report the three condition numbers for each problem. There is no strict ordering between these condition numbers (partly, again, because of the choice of norm), but there are constants c_1 and c_2 depending only on n such that

$$\mathrm{cond}_2(A, x) \leq c_1 \mathrm{cond}_2(A) \leq c_2 \kappa_2(A)$$

(see (1.10) and section 2).

The results are presented in Tables 4.1–4.6. The matrices A in Tables 4.1–4.3 are random matrices with geometrically distributed singular values $\sigma_i = \alpha^i$, generated usin the routine randsvd of [10]. In Table 4.4, $Ax = b$ is the same system used in Table 4.1 but with the fifth equation scaled by $2^{15} = 32768$. In Table 4.5 the system is the one used in Table 4.1 but with the eighth column of A scaled by 2^{15}. In Table 4.6, A is a Kahan matrix—an ill-conditioned upper trapezoidal matrix with rows of widely varying norm [7, p. 245], [10].

The key features in the results are as follows.

(1) The error bounds of the previous section are confirmed. Indeed for both the Q method and the SNE method the heuristic

$$\gamma_2(\widehat{x}) = \frac{\|\widehat{x} - x\|_2}{\|x\|_2} \approx \mathrm{cond}_2(A)u$$

predicts the error correct to within an order of magnitude in these examples.

(2) The independence of the forward errors on the row scaling of A is illustrated by Tables 4.1 and 4.4. However, column scaling can have an adverse effect, as shown in Table 4.5.

(3) The relative residuals confirm that the Q method is (almost) row-wise backward stable and that the SNE method is not even normwise backward

stable. The relative residuals for the SNE method exhibit dependence on $\mathrm{cond}_2(A)$ in these examples (dependence of the normwise residual on $\kappa_2(A)$ in the case of overdetermined systems is proven by Björck in [4, Th. 3.1]). Iterative refinement can produce a small relative residual, but can fail on very ill-conditioned problems, as in Table 4.3.

The condition numbers displayed in the tables can all be estimated cheaply given a QR factorization of A^T. For example, we show how to estimate $\mathrm{cond}_2(A,x)$. This differs by at most a factor $\sqrt{n}$ from $\mathrm{cond}_\infty(A,x)$. We consider only the first term of $\mathrm{cond}_\infty(A,x)$ in (2.3), as the second term can be treated similarly. As in [1], we can convert this norm of a vector into a norm of a matrix: with $g = |A^T||A^{+T}x|$ and $G = \mathrm{diag}(g_i)$, we have

$$\begin{aligned}
\|\,|I - A^+A|\cdot|A^T|\cdot|A^{+T}x|\,\|_\infty &= \|\,|I - A^+A|g\,\|_\infty \\
&= \|\,|I - A^+A|Ge\,\|_\infty = \|\,|I - A^+A|G\,\|_\infty \\
&= \|\,|(I - A^+A)G|\,\|_\infty \\
&= \|\,(I - A^+A)G\,\|_\infty.
\end{aligned}$$

The latter norm can be estimated by the method of [8] and [9, 11], which estimates $\|B\|_1$ given a means for forming matrix-vector products Bx and B^Ty. Forming these products for $B^T = (I - A^+A)G$ involves multiplying by G and Q or their transposes, and solving triangular systems with R and R^T.

Table 4.1: $A = \mathrm{randsvd}([10, 16], 1\mathrm{e}2)$

$\kappa_2(A) = 1\mathrm{e}2$, $\mathrm{cond}_2(A) = 8.63\mathrm{e}1$, $\mathrm{cond}_2(A, x) = 1.57\mathrm{e}2$

	$\rho^N(\widehat{y})$	$\rho^R(\widehat{y})$	$\rho^C(\widehat{y})$	$\gamma_2(\widehat{y})$
Q method	1.83e-8	9.88e-9	1.42e-7	2.01e-6
SNE	5.11e-7	2.79e-7	4.40e-6	4.97e-6
	1.52e-8	6.45e-9	9.64e-8	1.99e-6

Table 4.2: $A = \mathrm{randsvd}([10, 16], 1\mathrm{e}4)$

$\kappa_2(A) = 1\mathrm{e}4$, $\mathrm{cond}_2(A) = 5.43\mathrm{e}3$, $\mathrm{cond}_2(A, x) = 1.05\mathrm{e}4$

	$\rho^N(\widehat{y})$	$\rho^R(\widehat{y})$	$\rho^C(\widehat{y})$	$\gamma_2(\widehat{y})$
Q method	5.16e-9	6.84e-9	9.16e-8	1.29e-4
SNE	1.30e-5	1.56e-5	2.36e-4	2.30e-4
	4.29e-9	4.63e-9	8.01e-8	1.04e-4

Table 4.3: $A = \mathrm{randsvd}([10, 16], 1\mathrm{e}6)$

$\kappa_2(A) = 1\mathrm{e}6$, $\mathrm{cond}_2(A) = 4.30\mathrm{e}5$, $\mathrm{cond}_2(A, x) = 8.86\mathrm{e}5$

	$\rho^N(\widehat{y})$	$\rho^R(\widehat{y})$	$\rho^C(\widehat{y})$	$\gamma_2(\widehat{y})$
Q method	6.88e-9	5.78e-9	9.18e-8	6.50e-3
SNE	3.58e-3	1.69e-3	2.56e-2	2.47e-2
	5.17e-5	2.47e-5	3.74e-4	1.28e-2
	5.39e-6	2.62e-6	3.96e-5	1.11e-2
	2.05e-5	9.33e-6	1.41e-4	1.11e-2
	1.51e-5	6.94e-6	1.05e-4	1.27e-2

Table 4.4: $A = D\,\mathrm{randsvd}([10,16], 1\mathrm{e}2)$

$\kappa_2(A) = 1.63\mathrm{e}6$, $\mathrm{cond}_2(A) = 8.63\mathrm{e}1$, $\mathrm{cond}_2(A, x) = 1.57\mathrm{e}2$

	$\rho^N(\widehat{y})$	$\rho^R(\widehat{y})$	$\rho^C(\widehat{y})$	$\gamma_2(\widehat{y})$
Q method	9.24e-9	9.88e-9	1.42e-7	2.01e-6
SNE	9.26e-7	2.79e-7	4.40e-6	4.97e-6
	5.70e-9	6.45e-9	9.64e-8	1.99e-6

Table 4.5: $A = \mathrm{randsvd}([10,16], 1\mathrm{e}2)\,D$

$\kappa_2(A) = 1.37\mathrm{e}6$, $\mathrm{cond}_2(A) = 7.81\mathrm{e}5$, $\mathrm{cond}_2(A, x) = 1.35\mathrm{e}2$

	$\rho^N(\widehat{y})$	$\rho^R(\widehat{y})$	$\rho^C(\widehat{y})$	$\gamma_2(\widehat{y})$
Q method	2.42e-9	5.70e-9	2.95e-4	9.29e-3
SNE	4.23e-3	5.89e-3	9.98e-1	2.61e-2
	1.39e-5	1.93e-5	5.89e-1	1.75e-3
	4.24e-7	5.90e-7	4.34e-2	3.60e-5
	3.02e-9	4.20e-9	3.12e-4	1.29e-6

Table 4.6: $A = \mathrm{kahan}([10,16])$

$\kappa_2(A) = 6.29\mathrm{e}5$, $\mathrm{cond}_2(A) = 9.58\mathrm{e}0$, $\mathrm{cond}_2(A, x) = 1.02\mathrm{e}1$

	$\rho^N(\widehat{y})$	$\rho^R(\widehat{y})$	$\rho^C(\widehat{y})$	$\gamma_2(\widehat{y})$
Q method	1.22e-8	3.52e-9	4.99e-8	1.79e-7
SNE	8.00e-8	3.42e-8	3.27e-7	3.35e-7

5 Implications for Square Linear Systems

All the results in sections 2 and 3 are valid when $m = n$. Theorem 2.1 reduces to a straightforward generalization of a result in [17, Th. 2.1]. However, the error bound

$$\frac{\|\widehat{x} - x\|_\infty}{\|x\|_\infty} \le \mu_n \mathrm{cond}_\infty(A) u + O(u^2) \tag{5.1}$$

for the Q method is not a familiar one for square systems. (We have switched to the ∞-norm, which is the more usual choice for square systems). In fact, a bound of the form (5.1) holds also if we solve $Ax = b$ using an LU factorization (with partial pivoting) of A^T. Of course, when solving a square system $Ax = b$ it is more natural to employ an LU or QR factorization of A than of A^T. But if a factorization of A is used then no bound of the form (5.1) holds in general—the best we can say is that

$$\frac{\|\widehat{x} - x\|_\infty}{\|x\|_\infty} \le \mu_n \kappa_\infty(A) u + O(u^2). \tag{5.2}$$

We note, however, that there is a simple way to achieve a bound of the form (5.1) for LU and QR factorization of A: work with the scaled system $(DA)x = Db$ instead of $Ax = b$, where $B = DA$ has rows of unit 1-norm. This follows from (5.2) and the fact that $\kappa_\infty(B) = \mathrm{cond}_\infty(A)$. To verify the latter equality note that if $D^{-1} = \mathrm{diag}(|A|e)$, then

$$\begin{aligned}
\mathrm{cond}_\infty(A) &= \|\, |A^{-1}| \cdot |A| \,\|_\infty = \|\, |A^{-1}| \cdot |A|e \,\|_\infty = \|\, |A^{-1}|D^{-1}e \,\|_\infty \\
&= \|\, |A^{-1}|D^{-1} \,\|_\infty = \|\, |A^{-1}D^{-1}| \,\|_\infty = \|\, |(DA)^{-1}| \,\|_\infty \\
&= \|\, B^{-1} \,\|_\infty = \kappa_\infty(B).
\end{aligned}$$

It is interesting to compare this row equilibration strategy with fixed precision iterative refinement (FPIR). It is known that under suitable assumptions FPIR in conjunction with LU factorization with partial pivoting [1, 18] or QR factorization [12] leads to a computed $\widehat{y}$ such that $\omega^C(\widehat{y}) = O(u)$, that is, FPIR brings componentwise backward stability. ¿From an ∞-norm version of Theorem 2.1 we see that $\omega^C(\widehat{y}) \le u$ implies

$$\frac{\|\widehat{y} - x\|_\infty}{\|x\|_\infty} \le 2\,\mathrm{cond}_\infty(A,x)u + O(u^2),$$

where

$$\mathrm{cond}_\infty(A,x) = \frac{\|\,|A^{-1}|\cdot|A|\cdot|x|\,\|_\infty}{\|x\|_\infty}.$$

This is a stronger bound than (5.1) because $\mathrm{cond}_\infty(A,x) \le \mathrm{cond}_\infty(A)$ (with equality for $x = e$) and for some A and x, $\mathrm{cond}_\infty(A,x) \ll \mathrm{cond}_\infty(A)$ (see, for example, a 3×3 example of Hamming quoted in [17, p. 500]).

Skeel [17, 19] looks in detail at the possible benefits of row scaling for LU factorization. In [17, sec.4.2] he shows that for the scaling $D^{-1} = \mathrm{diag}(|A||x|)$ the forward error bound is proportional to $\mathrm{cond}_\infty(A,x)$; unfortunately, since x is unknown this "optimal" scaling is of little practical use. Row equilibration can be regarded as approximating $|x|$ by e in the optimal scaling.

To sum up, we regard row equilibration as a "quick and dirty" way to achieve a "cond-bounded" forward error—quick because the scaling is trivial to perform, and dirty because the forward error bound is independent of the right-hand side b and there is no guarantee that a small componentwise backward error will be achieved. In contrast, FPIR produces a small componentwise backward error and has a sharper forward error bound that depends on b (but FPIR may fail to converge).

We illustrate our observations with two numerical examples computed using MATLAB in simulated single precision, as in section 4. For odd $n = 2k+1$ let V_n be the Vandermonde matrix with (i,j) element $(-k + j - 1)^{i-1}$. We solved two systems $V_n x = b$ by both LU factorization with partial pivoting and QR factorization, in each case trying both FPIR and the row equilibration discussed above.

The two systems were chosen to illustrate two extreme cases. For the first problem, $V_9 e = b$ reported in Table 5.1, $\mathrm{cond}_\infty(A) = \mathrm{cond}_\infty(A,x) \approx$

$\frac{1}{359}\kappa_\infty(A)$ and row equilibration is about as effective as FPIR as measured by the size of the componentwise backward error and the relative error. For the second system, $V_{11}x = e$, $\mathrm{cond}_\infty(A,x) \approx \frac{1}{174}\mathrm{cond}_\infty(A) \ll \kappa_\infty(A)$ and FPIR achieves a significantly smaller componentwise backward error and relative error than row equilibration.

We also tried using a scaling obtained by perturbing the equilibrating transformation $D = \mathrm{diag}(|A|e)^{-1}$ to the nearest powers of 2, so as not to introduce rounding errors. This led to final errors sometimes larger and sometimes smaller than with D. In any case, from the point of view of the error bounds the rounding errors introduced by the scaling are easily seen to be insignificant.

Table 5.1: $A = V_9$, $x = e$

$\kappa_\infty(A) = 4.27\text{e}5$, $\text{cond}_\infty(A) = 1.19\text{e}3$, $\text{cond}_\infty(A,x) = 1.19\text{e}3$

	$\omega^N(\widehat{y})$	$\omega^R(\widehat{y})$	$\omega^C(\widehat{y})$	$\gamma_\infty(\widehat{y})$
LU with FPIR	2.11e-8	6.25e-7	3.13e-6	1.81e-3
	1.65e-8	1.65e-8	8.26e-8	1.79e-5
LU with equilibration	6.74e-9	1.91e-8	1.72e-7	2.38e-5
QR with FPIR	1.13e-8	7.47e-6	6.73e-5	3.85e-3
	3.51e-9	1.06e-8	8.28e-8	1.44e-5
QR with equilibration	2.3e-8	3.71e-8	3.34e-7	1.60e-4

Table 5.2: $A = V_{11}$, $b = e$

$\kappa_\infty(A) = 6.68\text{e}7$, $\text{cond}_\infty(A) = 9.17\text{e}3$, $\text{cond}_\infty(A,x) = 5.27\text{e}1$

	$\omega^N(\widehat{y})$	$\omega^R(\widehat{y})$	$\omega^C(\widehat{y})$	$\gamma_\infty(\widehat{y})$
LU with FPIR	2.18e-12	2.57e-7	4.82e-6	5.23e-5
	4.57e-12	1.53e-9	5.83e-8	6.83e-7
LU with equilibration	1.22e-10	9.96e-9	2.88e-6	6.24e-5
QR with FPIR	1.95e-11	1.64e-5	1.48e-4	3.59e-4
	4.48e-12	4.86e-9	9.96e-8	1.38e-6
QR with equilibration	3.75e-9	4.75e-9	5.83e-6	1.38e-5

References

[1] M. Arioli, J.W. Demmel and I.S. Duff, Solving sparse linear systems with sparse backward error, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 165–190.

[2] M. Arioli and A. Laratta, Error analysis of an algorithm for solving an underdetermined linear system, Numer. Math., 46 (1985), pp. 255–268.

[3] F.L. Bauer, Genauigkeitsfragen bei der Lösung linearer Gleichungssysteme, Z. Angew. Math. Mech., 46 (1966), pp. 409–421.

[4] Å. Björck, Stability analysis of the method of seminormal equations for linear least squares problems, Linear Algebra and Appl., 88/89 (1987), pp. 31–48.

[5] R.E. Cline and R.J. Plemmons, l_2-solutions to underdetermined systems, SIAM Review, 18 (1976), pp. 92–106.

[6] P.E. Gill and W. Murray, A numerically stable form of the simplex algorithm, Linear Algebra and Appl., 7 (1973), pp. 99–138.

[7] G.H. Golub and C.F. Van Loan, *Matrix Computations*, Second Edition, Johns Hopkins University Press, Baltimore, Maryland, 1989.

[8] W.W. Hager, Condition estimates, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 311–316.

[9] N.J. Higham, FORTRAN codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation (Algorithm 674), ACM Trans. Math. Soft., 14 (1988), pp. 381–396.

[10] N.J. Higham, A collection of test matrices in MATLAB, Technical Report 89-1025, Department of Computer Science, Cornell University, 1989; to appear in ACM Trans. Math. Soft.

[11] N.J. Higham, Experience with a matrix norm estimator, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 804–809.

[12] N.J. Higham, Iterative refinement enhances the stability of QR factorization methods for solving linear equations, Numerical Analysis Report No. 182, University of Manchester, England, 1990.

[13] C.L. Lawson and R.J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

[14] W. Oettli and W. Prager, Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides, Numer. Math., 6 (1964), pp. 405–409.

[15] C.C. Paige, An error analysis of a method for solving matrix equations, Math. Comp., 27 (1973), pp. 355–359.

[16] M.A. Saunders, Large-scale linear programming using the Cholesky factorization, Report CS 252, Computer Science Department, Stanford University, 1972.

[17] R.D. Skeel, Scaling for numerical stability in Gaussian elimination, J. Assoc. Comput. Mach., 26 (1979), pp. 494–526.

[18] R.D. Skeel, Iterative refinement implies numerical stability for Gaussian elimination, Math. Comp., 35 (1980), pp. 817–832.

[19] R.D. Skeel, Effect of equilibration on residual size for partial pivoting, SIAM J. Numer. Anal., 18 (1981), pp. 449–454.

[20] P.-Å. Wedin, Perturbation theory for pseudo-inverses, BIT, 13 (1973), pp. 217–232.