NORTH-HOLLAND

Hybrid Norms and Bounds for Overdetermined Linear Systems

Arnold Neumaier

Institut für Angewandte Mathematik
Universität Freiburg
D-7800 Freiburg, Germany

Submitted by Richard A. Brualdi

ABSTRACT

Assuming exact arithmetic, rigorous componentwise error bounds are given for the solution of a linear least squares problem when interpreted as the exact solution of a perturbed system with known bounds on the perturbations. It turns out that the quality of the bounds is much better when this perturbed system is assumed to be consistent. The sparse case is discussed, too.

1. INTRODUCTION

In this paper we consider the problem of assessing the quality of the least squares solution x^* of the problem

$$\|Ax - b\|_2 = \min! \qquad (b \in \mathbb{R}^n, \quad A \in \mathbb{R}^{m \times n} \text{ of rank } n \leq m),$$

assuming that bounds for the errors in A and b are known. Since the accuracy of the component x_i may be rather different, we are interested in componentwise bounds. Therefore it is reasonable to assume that the error in A is bounded separately for each column. More precisely, we assume that the following situation holds:

(A1) $A, \widetilde{A} \in \mathbb{R}^{m \times n}$ and $b, \widetilde{b} \in \mathbb{R}^n$ satisfy

$$\|\widetilde{A}_{\bullet k} - A_{\bullet k}\|_2 \leq c_k \quad (k = 1, \ldots, n), \qquad \|\widetilde{b} - b\|_2 \leq \beta.$$

(A2) $\widetilde{x} \in \mathbb{R}^n$ satisfies $\widetilde{A}\widetilde{x} = \widetilde{b}$, and x^* solves the least squares problem

$$\|Ax - b\|_2 = \min! \qquad \text{so that} \quad A^T A x^* = A^T b.$$

Here, as later, we denote the ith row of A by $A_{i\bullet}$ and the kth column of A by $A_{\bullet k}$.

LINEAR ALGEBRA AND ITS APPLICATIONS 216:257–265 (1995)

© Elsevier Science Inc., 1995
655 Avenue of the Americas, New York, NY 10010

0024-3795/95/$9.50
SSDI 0024-3795(93)00152-P

In applications, A and b might be measured approximations to unknown quantities $\widetilde{A}$, $\widetilde{b}$. Often, the accuracy of the measurements can be translated into bounds for the errors $\widetilde{A} - A$ and $\widetilde{b} - b$; we bound them columnwise by c^T and β. The required solution $\widetilde{x}$ then lies "near" the least squares solution x^*. We show that a bound for the error $\widetilde{x} - x^*$ can be found from a QR factorization of A (with orthogonal Q and upper triangular R). We also obtain bounds when (A2) is replaced by the weaker assumption

(A3) $\widetilde{x} \in \mathbb{R}^n$ satisfies $\|\widetilde{A}\widetilde{x} - \widetilde{b}\|_2 = \min!$, and x^* solves the least squares problem

$$\|Ax - b\|_2 = \min! \qquad \text{so that} \quad A^T A x^* = A^T b;$$

of course the resulting bounds will be weaker, too.

2. HYBRID NORMS

Componentwise bounds are most easily handled with the concept of hybrid norms.

PROPOSITION. *Let $A \in \mathbb{R}^{m \times n}$, and define the vector-valued hybrid norms $\nu_2 : \mathbb{R}^{m \times n} \to \mathbb{R}^m$ and $\nu_2^T : \mathbb{R}^{m \times n} \to (\mathbb{R}^n)^T$ by*

$$\nu_2(A) := \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_m \end{pmatrix}, \qquad \text{where} \quad \gamma_i := \|A_{i\bullet}\|_2,$$
$$\nu_2^T(A) := (c_1, \ldots, c_n), \quad \text{where} \quad c_k := \|A_{\bullet k}\|_2.$$

Then, for all $x \in \mathbb{R}^n$, we have

$$|Ax| \le \nu_2(A)\|x\|_2, \tag{1}$$
$$|x^T A| \le \|x\|_2 \nu_2^T(A), \tag{2}$$
$$\|Ax\|_2 \le \nu_2^T(A)|x|; \tag{3}$$

here absolute values and inequalities for vectors are taken componentwise.

Moreover,

$$\|A\|_2 \le \|\nu_2(A)\| = \|\nu_2^T(A)\| = \sqrt{\mathrm{tr} A^T A} \ (= \|A\|_F) \tag{4}$$

and

$$\|AB\|_2 \le \|\nu_2^T(A)|B|\|_2. \tag{5}$$

PROOF. By the Cauchy-Schwarz inequality, (1) follows from

$$|Ax|_i = |A_{i\bullet}x| \le \|A_{i\bullet}\|_2 \|x\|_2 = \gamma_i \|x\|_2,$$

and (2) from

$$|x^T A|_k = |x^T A_{\bullet k}| \le \|x\|_2 \|A_{\bullet k}\|_2 = \|x\|_2 c_k.$$

(3) follows from

$$\|Ax\|_2 = \left\| \sum_k A_{\bullet k} x_k \right\|_2 \le \sum_k \|A_{\bullet k}\|_2 |x_k| = \sum_k c_k |x_k| = \nu_2^T(A)|x|.$$

Taking norms in (1) gives (4), and inserting Bx for x into (3) gives (5), using the Cauchy-Schwarz inequality.

In view of the applications to least squares problems, we shall only consider the 2-norm. However, similar statements hold for other monotone norms. For example, we have for the maximum norm

$$|Ax| \le \nu_\infty(A)\|x\|_\infty, \qquad \|Ax\|_\infty \le \nu_\infty^T(A)|x|,$$

where

$$\nu_\infty(A) := (\beta_1, \ldots, \beta_m)^T \qquad \beta_i := \|A_{i\bullet}\|_1,$$
$$\nu_\infty^T(A) := (b_1, \ldots, b_n), \qquad b_k := \|A_{\bullet k}\|_\infty.$$

3. MAIN RESULTS

In this section we prove two error estimates, assuming (A1) and either (A2) or (A3).

THEOREM 1. *Assume that A has rank n and QR factorization $A = QR$. Define*

$$\rho := \|Ax^* - b|_2, \qquad \sigma := \beta + c^T|x^*|, \qquad f := \nu_2(R^{-1}).$$

(i) *If* (A1) *and* (A2) *hold and $c^T f < 1$, then*

$$|\widetilde{x} - x^*| \le \bar{\gamma} f, \qquad \text{where} \quad \bar{\gamma} := \frac{\sigma c^T f + \sqrt{\sigma^2 - \rho^2[1 - (c^T f)^2]}}{1 - (c^T f)^2}$$

(ii) *If $\sigma^2 < \rho^2[1 - (c^T f)^2]$, then assumptions* (A1) *and* (A2) *are inconsistent.*

PROOF. Put $\delta := \tilde{x} - x^*$, $r := Ax^* - b$. We first show the relation

$$\|r + A\delta\|_2^2 = \|r\|_2^2 + \|R\delta\|_2^2. \tag{6}$$

Indeed, we have $A^T r = A^T A x^* - A^T b = 0$ and $A^T A = R^T Q^T Q R = R^T R$, so that (6) follows from

$$\begin{aligned} \|r + A\delta\|_2^2 &= (r + A\delta)^T (r + A\delta) = r^T r + \delta^T A^T r + (A^T r)^T \delta + \delta^T A^T A \delta \\ &= r^T r + \delta^T R^T R \delta = \|r\|_2^2 + \|R\delta\|_2^2. \end{aligned}$$

Now, with $\gamma := \|R\delta\|_2$, we have

$$|\delta| = |R^{-1} R\delta| \le \nu_2(R^{-1}) \|R\delta\|_2 = \gamma f, \tag{7}$$

and the vector

$$d := \tilde{b} - b + (A - \tilde{A})\tilde{x} \tag{8}$$

satisfies

$$\begin{aligned} \|d\|_2 &= \|\tilde{b} - b + (A - \tilde{A})\tilde{x}\|_2 \\ &\le \|\tilde{b} - b\|_2 + \nu_2^T (A - \tilde{A}) |\tilde{x}| \\ &\le \beta + c^T |x^* + \delta| \le \beta + c^T |x^*| + c^T |\delta|; \end{aligned}$$

hence

$$\|d\|_2 \le \sigma + \gamma c^T f. \tag{9}$$

Since $\tilde{A}\tilde{x} = \tilde{b}$, we find from (6), (8), and (9) the relation

$$\rho^2 + \gamma^2 = \|r + A\delta\|_2^2 = \|Ax^* - b + A(\tilde{x} - x^*)\|_2^2 = \|d\|_2^2 \le (\sigma + \gamma c^T f)^2.$$

Solving this quadratic inequality for γ, we find $\gamma \le \bar{\gamma}$ when $\sigma^2 \ge \rho^2 [1 - (c^T f)^2]$, and a contradiction otherwise. The assertion hence follows from (7).

THEOREM 2. *Assume that A has rank n and QR factorization $A = QR$. Define*

$$\rho := \|Ax^* - b\|_2, \qquad \sigma := \beta + c^T |x^*|, \qquad f := \nu_2(R^{-1}).$$

If (A1) *and* (A3) *hold and $c^T f < 1$, then*

$$|\tilde{x} - x^*| \le \hat{\gamma} f, \qquad \text{where} \quad \hat{\gamma} := \frac{\sigma + \frac{\omega\tau^2}{2} + \tau\sqrt{\rho^2 + \beta\rho + \omega\sigma + \frac{\omega^2\tau^2}{4}}}{1 - c^T f}$$

with

$$\omega = \frac{\|b\|_2 + \beta}{1 - c^T f}, \qquad \tau = \left\| |c^T |R^{-1}| \right\|_2.$$

PROOF. As in the previous proof, (4)–(7) hold. But the residual $\tilde{r} := \tilde{A}\tilde{x} - \tilde{b}$ now can be nonzero, and we only know $A^T r = \tilde{A}^T \tilde{r} = 0$. To motivate the argument, assume first that $\tilde{A} = A$. Then $A\delta = A\tilde{x} - Ax^* = \tilde{r} - r + \tilde{b} - b$; hence

$$R^T R\delta = A^T A\delta = A^T(\tilde{r} - r) + A^T(\tilde{b} - b) = A^T(\tilde{b} - b) = R^T Q^T(\tilde{b} - b),$$

whence

$$\gamma := \|R\delta\|_2 = \left\|Q^T(\tilde{b} - b)\right\|_2 \le \|Q\|_2 \|\tilde{b} - b\|_2 \le \|\tilde{b} - b\|_2 \le \beta.$$

If $\tilde{A} \ne A$, we must refine the argument to take account of the error $\tilde{A} - A$. Now $A\delta = A(\tilde{x} - x^*) = (A - \tilde{A})\tilde{x} + \tilde{A}\tilde{x} - Ax^* = (A - \tilde{A})\tilde{x} + \tilde{r} + \tilde{b} - r - b$; hence by (8)

$$A\delta = d + \tilde{r} - r. \tag{10}$$

This implies

$$R^T R\delta = A^T A\delta = A^T(d + \tilde{r} - r) = A^T d + A^T \tilde{r} = R^T Q^T d - (\tilde{A} - A)^T \tilde{r},$$

where

$$R\delta = Q^T d - R^{-T}(\tilde{A} - A)^T \tilde{r} \tag{11}$$

Thus $\gamma = \|Rd\|_2 \le \|Q\|_2 \|d\|_2 + \|(\tilde{A} - A)R^{-1}\|_2 \|\tilde{r}\|_2$. Since Q is orthogonal and (5) implies $\|(\tilde{A} - A)R^{-1}\|_2 \le \|c^T |R^{-1}|\|_2 = \tau$, we find

$$\gamma \le \|d\|_2 + \tau \|\tilde{r}\|_2 =: \gamma_0. \tag{12}$$

To bound the unknown residual norm in (12) we note that

$$\|r\|_2^2 = r^T r = x^{*T} A^T r - b^T r = -b^T r$$

and similarly $\|\tilde{r}\|_2^2 = \tilde{b}^T \tilde{r}$, so that

$$\|\tilde{r}\|_2^2 = -\tilde{b}^T \tilde{r} = \|r\|_2^2 + (b - \tilde{b})^T r + \tilde{b}^T (r - \tilde{r}) \le \rho^2 + \beta\rho + \|\tilde{b}\|_2 \|r - \tilde{r}\|_2.$$

Now $\|\widetilde{b}\|_2 \le \|b\|_2 + \|\widetilde{b} - b\|_2 \le \|b\|_2 + \beta$ and, by (10) and (11),

$$\begin{aligned} \|r - \widetilde{r}\|_2 &= \|d - A\delta\|_2 = \|d - QR\delta\|_2 \\ &= \left\|(I - QQ^T)d + QR^{-T}(\widetilde{A} - A)^T\widetilde{r}\right\|_2 \\ &\le \left\|I - QQ^T\right\|_2 \|d\|_2 + \|Q\|_2 \left\|(\widetilde{A} - A)R^{-1}\right\|_2 \|\widetilde{r}\|_2 \\ &\le \|d\|_2 + \tau\|\widetilde{r}\|_2 = \gamma_0. \end{aligned}$$

Therefore

$$\|\widetilde{r}\|_2^2 \le \rho^2 + \beta\rho + \big(\|b\|_2 + \beta\big)\gamma_0. \tag{13}$$

Now (12) and (9) give $\gamma_0 \le \sigma + \gamma_0 c^T f + \tau\|\widetilde{r}\|_2$, so that

$$\gamma_0 \le \frac{\sigma + \tau\|\widetilde{r}\|_2}{1 - c^T f}. \tag{14}$$

Inserting this into (13) yields

$$\|\widetilde{r}\|_2^2 \le \rho^2 + \beta\rho + \omega\big(\sigma + \tau\|\widetilde{r}\|_2\big),$$

and the solution of this quadratic inequality is

$$\|\widetilde{r}\|_2 \le \frac{\omega\tau}{2} + \sqrt{\rho^2 + \beta\rho + \omega\sigma + \frac{\omega^2\tau^2}{4}}.$$

If we insert this into (14), we finally get the desired conclusion $\gamma_0 \le \widehat{\gamma}$ and hence $\gamma \le \widehat{\gamma}$ by (12). ∎

4. DISCUSSION

We now comment on the bounds produced in the previous section.

1. The condition $c^T f < 1$ requires that the errors in A be sufficiently small (roughly speaking, less than the inverse of the absolutely largest entry of R^{-1}) and guarantees that all matrices $\widetilde{A}$ with (A1) have rank n, too.

2. The theorem assumes that x^* is the exact solution of the least squares problem. In the presence of rounding errors, the bound $|\widetilde{x} - x^*| \le \overline{\gamma} f$ (with the *computed* x^*) is only approximately valid, and it may be too optimistic when the input data are rather accurate. However, an acceptable estimate for $\overline{\gamma}$ is obtained if we determine the error quantities c and β instead from

$$4n\varepsilon\nu_2^T(A) + \nu_2^T(\widetilde{A} - A) \le c^T, \qquad \varepsilon\|b\|_2 + \|\widetilde{b} - b\|_2 \le \beta,$$

where ε is the relative machine accuracy. (A rigorous analysis of rounding errors is much more difficult, and is not considered here.)

3. If $c^T f \ll 1$, then $\overline{\gamma} \approx \sqrt{\sigma^2 - \rho^2} \le \sigma$ is of the order of the input errors; the components f_i of f therefore measure the maximal error magnification in the components x_i of the solution. It can be shown that, for fixed i, the bound $|\widetilde{x}_i - x_i^*| \le \overline{\gamma} f_i$ can be attained under assumptions (A1) and (A2) up to a factor $1 + O(c^T f + \rho)$; hence the f_i can be regarded as componentwise condition numbers of the least squares problem.

4. Other bounds for the norm of the error in least squares problems are discussed in [4, 7–9]. Rigorous bounds using interval analysis can be found in Gay [2], Rump [6]. Approximate componentwise bounds assuming (A1) and (A2) have been obtained by Manteuffel [5] by neglecting higher order error terms. In our notation, Manteuffel obtains the bound

$$|\widetilde{x} - x^*| \le f\sigma + O(\varepsilon^2) \qquad \text{when} \quad \beta,\, c + O(\varepsilon).$$

In the formula of Theorem 1, the higher order terms are covered by the terms $c^T f$ in the expression for $\overline{\gamma}$; ignoring them still gives the bound

$$|\widetilde{x} - x^*| \le f\sqrt{\sigma^2 - \rho^2} + O(\varepsilon^2), \tag{15}$$

an improvement over Manteuffel's result.

5. If we neglect higher order terms in Theorem 2, we find similarly

$$|\widetilde{x} - x^*| \le f(\sigma + \tau\rho) + O(\varepsilon^2),$$

which, especially for large errors in the matrix, is inferior to the result (14) obtained from the assumption of a nearby feasible solution. It is a componentwise version of a bound in Wedin [9]. A comparison with the previous remark 4 shows that the bounds obained from assumption (A2) are much better than those obtained from the weaker assumption (A3).

6. The calculation of f requires R^{-1}; hence $n^3/6 + O(n^2)$ multiplications are needed. But for $m \gg n$ this extra effort is small compared with the $(m - n/3)n^2 + O(mn)$ multiplications required for the QR factorization. A (sometimes weak) upper bound for f can be computed in $O(n^2)$ operations; see Higham [3]. For sparse least squares problems (see Björck [1, Chapter III] and references there) one often uses a sparse QR factorization. In this case it is not advisable to compute R^{-1} explicitly, since nearly all sparsity is lost. Fortunately, there is a way of computing $f = \nu_2(R^{-1})$, without knowing R^{-1}. Indeed, the definition of the hybrid norm

implies that f_i^2 is just the ith diagonal entry of the matrix $R^{-1}R^{-T} = (R^TR)^{-1}$. Now it is possible to compute the entries of $(R^TR)^{-1}$ within the envelope of $R+R^T$ in $O(m^2n)$ operations, where m is the average half-band width; see e.g. Björck [1, p. 539]. In particular, this yields the required diagonal entries $f_i = \sqrt{(R^TR)_{ii}^{-1}}$.

7. By the same trick we can compute cheaply the Frobenius norm

$$\|A^+\|_F = \sqrt{\sum_{i,k}(A^+)_{ik}^2}$$

of the pseudoinverse $A^+ = R^{-1}Q^T$ of a sparse matrix A of full rank. Indeed,

$$\|A^+\|_F^2 = \operatorname{tr} A^+(A^+)^T = \operatorname{tr} R^{-1}Q^TQR^{-T} = \operatorname{tr} R^{-1}R^{-T} = \operatorname{tr}(R^TR)^{-1},$$

which is available from the envelope part of the inverse. In view of the well-known inequalities

$$\begin{aligned}
\frac{1}{\sqrt{n}}\|A^+\|_F &\le \|A^+\|_\infty \le \sqrt{m}\|A^+\|_F,\\
\frac{1}{\sqrt{m}}\|A^+\|_F &\le \|A^+\|_1 \le \sqrt{n}\|A^+\|_F,\\
\frac{1}{\sqrt{n}}\|A^+\|_F &\le \|A^+\|_2 \le \|A^+\|_F,
\end{aligned}$$

this gives also bounds for other common norms.

REFERENCES

1 Å. Björck, Least squares methods, in *Handbook of Numerical Analysis*, Vol. I (P. G. Ciarlet and J. L. Lions, Eds.), Elsevier, Amsterdam, 1990.
2 D. M. Gay, Interval Least squares—a diagnostic tool, in *Reliability in Computing* (R. E. Moore, Ed.), Academic London, 1988, pp. 183–206.
3 N. J. Higham, A survey of condition number estimation for triangular matrices, *SIAM Rev.* 29:575–596 (1987).
4 C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, N.J., 1974.
5 T. A. Manteuffel, An interval analysis approach to rank determination in linear least squares problems, *SIAM J. Sci. Statist. Comput.* 2:335–348 (1981).
6 S. M. Rump, Solving algebraic problems with high accuracy, in *A New Approach to Scientific Computation*, (U. W. Kulisch and W. L. Miranker, Eds.), Academic, New York, 1983, pp. 51–120.

7 G. W. Stewart, *Introduction to Matrix Computations*, Academic, New York, 1973.
8 G. W. Stewart, On the perturbation of pseudo-inverses, projections, and linear least squares, *SIAM Rev.* 19:634–662 (1977).
9 P.-Å. Wedin, Perturbation theory for pseudo-inverses, *BIT* 13:217–232 (1973).

Received 7 April 1992; final manuscript accepted 28 April 1993