submitted for publication March 20, 2013, accepted for
publication in Numerical Algorithms June 25, 2013

Improved componentwise verified error bounds for least squares problems and underdetermined linear systems

Siegfried M. Rump

the date of receipt and acceptance should be inserted later

Abstract Recently Miyajima presented algorithms to compute componentwise verified error bounds for the solution of full-rank least squares problems and underdetermined linear systems. In this paper we derive simpler and improved componentwise error bounds which are based on equalities for the error of a given approximate solution. Equalities are not improvable, and the expressions are formulated in a way that direct evaluation yields componentwise and rigorous estimates of good quality. The computed bounds are correct in a mathematical sense covering all sources of errors, in particular rounding errors. Numerical results show a gain in accuracy compared to previous results.

Mathematics Subject Classification (2000) 65F20, 65G20

Keywords Least squares problems, underdetermined linear systems, INTLAB, componentwise error estimates, normal equations, extra-precise residual evaluation.

1 Introduction and notation

For the solution of least squares problems and underdetermined linear systems a number of (normwise) backward stable algorithms are available [6, 7], which are usually based on a QR-decomposition of the matrix. Although numerical approximations are usually reliable, it seems desirable to provide rigorous error bounds, taking into account all errors, in particular rounding errors.

Such mathematically rigorous error bounds are mandatory in so-called computer-assisted proofs, where parts of a proof depend on the numerical solution of certain problems [5]. To maintain mathematical rigor, a numerical solution is accompanied by mathematically correct error bounds.

Famous examples of computer-assisted proofs are Tucker's paper [17], who was awarded the 2004 EMS prize by the European Mathematical Society for "giving a rigorous proof that the Lorenz attractor exists

S. M. Rump
Institute for Reliable Computing, Hamburg University of Technology, Schwarzenbergstraße 95,
Hamburg 21071, Germany,
and Visiting Professor at Waseda University, Faculty of Science and Engineering, 3–4–1 Okubo, Shinjuku-ku,
Tokyo 169–8555, Japan.
E-mail: rump@tuhh.de

for the parameter values provided by Lorenz. This was a long standing challenge to the dynamical system community, and was included by Smale in his list of problems for the new millennium. The proof uses computer estimates with rigorous bounds based on higher dimensional interval arithmetics."

As another example Sahinidis and Tawaralani [15] received the 2006 Beale-Orchard-Hays Prize for their package BARON which (citation) "incorporates techniques from automatic differentiation, interval arithmetic, and other areas to yield an automatic, modular, and relatively efficient solver for the very difficult area of global optimization".

Let $A \in \mathbb{K}^{m\times n}$, $b \in \mathbb{K}^m$, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. If A is rank-deficient, then the pseudoinverse does not depend continuously on the matrix data. This case is outside the scope of our methods because rigorous bounds are computed utilizing the speed of finite precision floating-point operations. For the moment we assume A to have full rank. Note, however, that this fact will be verified *a posteriori* by our methods. If this verification fails, then no bounds are computed. Thus computed bounds are always correct.

The 2-norm solution of the least squares problem is A^+b with A^+ denoting the Moore-Penrose pseudoinverse. Similarly, for $A \in \mathbb{K}^{n\times m}$, $b \in \mathbb{K}^n$, the minimum of $\|x\|_2$ subject to $Ax = b$ is achieved for $x = A^+b$. To avoid confusion, we specify rectangular matrices always such that $m \geq n$.

Mathematically, the least squares problem can be solved by an augmented linear system[1]

$$\begin{pmatrix} A & -I \\ \mathbf{0} & A^H \end{pmatrix} \begin{pmatrix} x \\ w \end{pmatrix} = \begin{pmatrix} b \\ \mathbf{0} \end{pmatrix} , \tag{1.1}$$

where I denotes the identity matrix and $\mathbf{0}$ denotes the zero matrix (vector) of proper dimension, respectively. Assume in the following that $\|A\|$ is of the order 1 for some norm. Then the condition number of the matrix in (1.1) is of the order $\operatorname{cond}(A)^2$. As shown by Björck [2] it can be reduced to about $\operatorname{cond}(A)$ by scaling $-I$; however, this is not used in the following.

For the least squares problem the solution vector of (1.1) satisfies $A^H w = \mathbf{0}$ and $x = A^+b$. Our verification methods are based on an economy-size QR-decomposition of A, that is $A = QR$ for unitary $Q \in \mathbb{K}^{m\times n}$ and triangular $R \in \mathbb{K}^{n\times n}$. Assume an approximate inverse S of (an approximate factor) R is given together with approximations $\widetilde{x}$ of A^+b and $\widetilde{w}$ near the kernel of A^H, i.e. $\widetilde{x}, \widetilde{w}$ are approximate solutions of (1.1). Define $X := AS$, so that based on our assumptions X can be expected to be not too far from orthogonality. Suppose $\|I - X^T X\|_p \leq \alpha < 1$ for some $p \in \{1, 2, \infty\}$. Then A has full rank, and bounds for $A^+b - \widetilde{x}$ based on $S, \widetilde{x}, \widetilde{w}$ can be computed. All remarks apply, *mutatis mutatandis*, to underdetermined linear systems.

Previously bounds for the error $A^+b - \widetilde{x}$ were derived by a sequence of estimates [10,14,11]. All those estimates are solely based on the approximations $S, \widetilde{x}, \widetilde{w}$. For example, (4.8) in [14] states[2]

$$\|A^+b - \widetilde{x}\|_p \leq \|SX^T \rho_{\widetilde{x}}\|_p + \|SS^T \rho_{\widetilde{w}}\|_p + \frac{\alpha\|S\|_p}{1-\alpha} \cdot \left(\|X^T \rho_{\widetilde{x}}\|_p + \|S^T \rho_{\widetilde{w}}\|_p\right) \quad \text{for } p \in \{1, 2, \infty\} , \tag{1.2}$$

where $\rho_{\widetilde{x}} := b - A\widetilde{x} + \widetilde{w}$ and $\rho_{\widetilde{w}} := A^T \widetilde{w}$. Based on his paper [10], Miyajima gave in [11] the following componentwise error estimate:

$$|A^+b - \widetilde{x}| \leq |SS^T(A^T \varrho_{\widetilde{x}} - \varrho_{\widetilde{w}})| + \frac{\|S^T(A^T \varrho_{\widetilde{x}} - \varrho_{\widetilde{w}}\|_\infty}{1-\alpha} |S||I - X^T X|\mathbf{e} , \tag{1.3}$$

[1] Sometimes (e.g. in [1,4]) the symmetric version of (1.1), obtained by interchanging the column blocks in the matrix, is used. However, this may lead to less accurate results, see the appendix.

[2] For a moment we restrict the attention to real problems.

where $\mathbf{e}$ denotes the vector of $1's$ of proper dimension, and comparison and absolute values are to be understood componentwise. Similar estimates are given for underdetermined systems, namely (3.8) in [14] states the normwise bound

$$\|A^+b-\widetilde{x}\|_p \le \sqrt{m}\|\rho_{\widetilde{w}}\|_p + \|Y^TS\rho_{\widetilde{x}}\|_p + \frac{\alpha\|Y^T\|_p}{1-\alpha}\;\|S\rho_{\widetilde{x}}\|_p \quad \text{for } p\in\{1,\infty\}, \tag{1.4}$$

and Miyajima [11] proves the componentwise bound

$$|A^+b-\widetilde{x}| \le \|\varrho_{\widetilde{w}}\|_2\mathbf{e} + |Y^TS\varrho_{\widetilde{x}}|\frac{\|S\varrho_{\widetilde{x}}\|_\infty}{1-\alpha}|Y^T|\;|E|\mathbf{e} \tag{1.5}$$

using $\rho_{\widetilde{w}} := \widetilde{x} - A^T\widetilde{w}$, $\rho_{\widetilde{x}} := A\widetilde{x} - b$, an approximate inverse S of R^T for an approximate decomposition $A^T \approx QR$ and $Y := SA$.

In this paper we derive simple expressions *equal to* $A^+b-\widetilde{x}$, also solely based on $S, \widetilde{x}, \widetilde{w}$. Those expressions are formulated in a way that componentwise and rigorous estimates of good quality can be computed. Note that mathematically $S, \widetilde{x}, \widetilde{w}$ are arbitrary quantities (of proper dimension); however, if they are of poor quality, then the bounds are of poor quality as well or no bounds may be computed at all.

For a matrix $M \in \mathbb{K}^{m\times n}$ and $1 \le i \le m$ denote by $M_{i*} \in \mathbb{K}^n$ the i-th row of M. For $1 \le p \le \infty$, we define the vector of row-wise p-norms of M by

$$v := \|M\|_p^{vec} \in \mathbb{R}^m \quad \text{by} \quad v_i := \|M_{i*}\|_p \text{ for } 1 \le i \le m\ . \tag{1.6}$$

The following application of Hölder's inequality is useful to obtain componentwise error bounds.

Lemma 1.1 *Let $M \in \mathbb{K}^{m\times n}$, $z \in \mathbb{K}^n$, and $1 \le p, q \le \infty$ with $1/p + 1/q = 1$. Then*

$$|Mz| \le \|z\|_p \cdot \|M\|_q^{vec}\ , \quad \textit{in particular} \quad |Mz| \le \|z\|_\infty \cdot |M|\mathbf{e}\ .$$

The most common choices for practical purposes are $p \in \{1, 2, \infty\}$.

Let $E \in \mathbb{K}^{n\times n}$ with $\|E\|_p \le \alpha < 1$ be given. Then it is well-known that $I - E$ is nonsingular,

$$\|(I-E)^{-1}\|_p \le \frac{1}{1-\alpha} \tag{1.7}$$

and

$$(I-E)^{-1} = I + (I-E)^{-1}E\ . \tag{1.8}$$

2 Main results

We begin with the componentwise error bounds for least squares problems.

Theorem 2.1 *Let $A \in \mathbb{K}^{m\times n}$, $b \in \mathbb{K}^m$, $S \in \mathbb{K}^{n\times n}$ with $m \ge n$ be given. Define $X := AS \in \mathbb{K}^{m\times n}$ and $E := I - X^HX$, and suppose $\|E\|_\infty \le \alpha < 1$. Let $\widetilde{x} \in \mathbb{K}^n$ and $\widetilde{w} \in \mathbb{K}^m$ be given and define*

$$\varrho_{\widetilde{x}} := b - A\widetilde{x} + \widetilde{w} \quad \textit{and} \quad \varrho_{\widetilde{w}} := A^H\widetilde{w} \quad \textit{and} \quad \delta := X^H\varrho_{\widetilde{x}} - S^H\varrho_{\widetilde{w}}\ . \tag{2.1}$$

Then

$$A^+b - \widetilde{x} = S(I-E)^{-1}\delta\ . \tag{2.2}$$

Therefore

$$|A^+b-\widetilde{x}| \le \frac{\|\delta\|_\infty}{1-\alpha}\cdot|S|\mathbf{e} \qquad \textit{and} \qquad |A^+b-\widetilde{x}| \le \frac{\|\delta\|_2}{1-\alpha}\cdot\|S\|_2^{vec}\ , \tag{2.3}$$

as well as

$$|A^+b-\widetilde{x}-S\delta| \le \frac{\|E\delta\|_\infty}{1-\alpha}\cdot|S|\mathbf{e} \qquad \textit{and} \qquad |A^+b-\widetilde{x}-S\delta| \le \frac{\|E\delta\|_2}{1-\alpha}\cdot\|S\|_2^{vec}\ . \tag{2.4}$$

Proof. It is well-known that $\|I-X^HX\|_\infty<1$ implies that X, and by $X=AS$ also A and X have full rank. Using $A^+A=I_n$, $A^+(A^+)^HA^H=A^+$, $A^+=SX^+$, $X^+=(X^HX)^{-1}X^H$ and $(X^HX)^{-1}=(I-E)^{-1}$ yields

$$\begin{aligned} A^+b-\widetilde{x} &= A^+\varrho_{\widetilde{x}} - A^+(A^+)^HA^H\widetilde{w} \\ &= SX^+\varrho_{\widetilde{x}} - SX^+(X^+)^HS^H\varrho_{\widetilde{w}} \\ &= S(X^HX)^{-1}X^H\varrho_{\widetilde{x}} - S(X^HX)^{-1}S^H\varrho_{\widetilde{w}} \\ &= S(I-E)^{-1}\delta\ . \end{aligned} \tag{2.5}$$

Applying (1.7), Lemma 1.1 and (1.8) prove the left estimates in (2.3) and (2.4). Furthermore

$$\|(I-E)^{-1}\|_2 \le \frac{1}{1-\|E\|_2} \le \frac{1}{1-\alpha} \tag{2.6}$$

using $\|E\|_2 \le \sqrt{\|E\|_1\|E\|_\infty} = \|E\|_\infty$ and $E^H=E$ prove the right estimates and thus the result. □

Note that (2.2) states an equality for the error $A^+b-\widetilde{x}$ and is thus not improvable. The only overestimation in (2.3) and (2.4) is introduced by the application of (1.7), Lemma 1.1 and (1.8). Practical experience suggest that this overestimation is not too large.

Also note that the quantities $X, \varrho_{\widetilde{x}}, \varrho_{\widetilde{w}}$ and δ are computed using $S, \widetilde{x}$ and $\widetilde{w}$. Thus all estimates in Theorem 2.1, as those in [10,14,11], are solely based on $S, \widetilde{x}, \widetilde{w}$.

Practical examples suggest that the leftmost bounds are usually the best ones, in particular better than using Lemma 1.1 with $p=1$ and $q=\infty$. Once the left bound in (2.3) or (2.4) is computed, the additional effort to compute the right bound is marginal. In any case the computing time is small compared to the QR-decomposition. So it seems advisable to compute both bounds in (2.3) or (2.4) and to take the componentwise minimum.

For S being an approximate inverse of R, the condition numbers of A and S can be expected to be of the same order. Therefore the quality of (2.2) depends mainly on $SS^H\varrho_{\widetilde{w}}$, implicitly included in $S(I-E)^{-1}\delta$. This seems unavoidable. But the condition number of SS^H is of the order $\mathrm{cond}(A)^2$. Thus in double precision and $\mathrm{cond}(A)$ beyond 10^8 bounds of good quality are only possible if $\varrho_{\widetilde{w}}$ is very small. This, in turn, is only achievable by representing the approximation $\widetilde{w}$ in two terms $\widetilde{w}_1+\widetilde{w}_2$. Similarly, for underdetermined linear systems the approximation $\widetilde{x}$ is represented in two terms $\widetilde{x}_1+\widetilde{x}_2$. For good bounds, both $\widetilde{x}$ and $\widetilde{w}$ should be improved by some residual iteration ensuring that both residuals $\varrho_{\widetilde{x}}$ and $\varrho_{\widetilde{w}}$ are very small, see the next section.

The second estimate (2.4) can be interpreted as improving the approximation $\widetilde{x}$ by some residual iteration but leaving the approximation $\widetilde{x}$ and correction $S\delta$ in separate parts. The method was introduced in [12] and became later [16] known as "staggered correction". Both is only meaningful if an accurate dot product is available. Note that the first summand in Miyajima's bound (1.3) is equal to $|S\delta|$, but the formulation is unfortunate for numerical evaluation. Moreover, our equality (2.2) allows to use the term $S\delta$ without absolute value in the left of (2.4).

A sample Matlab/INTLAB code to compute the bounds (1.2), (1.3), (2.3) and (2.4) for least squares problems is given in the appendix. For underdetermined linear systems we proceed similarly.

Note that the fact that A has full rank is not assumed a priori, but follows from $\alpha < 1$. Despite this, there is no further assumption, in particular not on the approximations $S, \widetilde{x}, \widetilde{w}$. Thus one might be inclined to define $\widetilde{w} := A\widetilde{x} - b$ so that $\varrho_{\widetilde{x}} = 0$. This is the best choice if $\widetilde{x}$ is equal to the solution A^+b. Otherwise, however, a good approximation $\widetilde{x}$ of A^+b does not ensure that $A\widetilde{x} - b$ is near the kernel of A^H. Similarly, one might set $\widetilde{w} := 0$ implying $\varrho_{\widetilde{w}} = 0$. However, then $\varrho_{\widetilde{x}} = b - A\widetilde{x}$ is in general not small.

Componentwise error estimates for underdetermined linear systems are established similar to Theorem 2.1.

Theorem 2.2 *Let $A \in \mathbb{K}^{n\times m}$, $b \in \mathbb{K}^n$, $S \in \mathbb{K}^{n\times n}$ with $m \ge n$ be given. Define $Y := SA \in \mathbb{K}^{n\times m}$ and $E := I - YY^H$, and suppose $\|E\|_\infty \le \alpha < 1$. Let $\widetilde{x} \in \mathbb{K}^m$ and $\widetilde{w} \in \mathbb{K}^n$ be given and define*

$$\varrho_{\widetilde{x}} := b - A\widetilde{x} \quad \text{and} \quad \varrho_{\widetilde{w}} := A^H\widetilde{w} - \widetilde{x} \quad \text{and} \quad \delta := S\varrho_{\widetilde{x}} - Y\varrho_{\widetilde{w}} \ . \tag{2.7}$$

Then

$$A^+b - \widetilde{x} - \varrho_{\widetilde{w}} = Y^H(I-E)^{-1}\delta \tag{2.8}$$

as well as

$$|A^+b - \widetilde{x} - \varrho_{\widetilde{w}}| \le \min\Big\{ \frac{\|\delta\|_\infty}{1-\alpha}\cdot|Y^H|\mathbf{e}\,,\ \frac{\|\delta\|_2}{1-\alpha}\cdot\|Y^H\|_2^{vec} \Big\}\ , \tag{2.9}$$

where the minimum is to be understood componentwise. Moreover,

$$|A^+b - \widetilde{x} - \varrho_{\widetilde{w}} - Y^H\delta| \le \min\Big\{ \frac{\|E\delta\|_\infty}{1-\alpha}\cdot|Y^H|\mathbf{e}\,,\ \frac{\|E\delta\|_2}{1-\alpha}\cdot\|Y^H\|_2^{vec} \Big\}\ . \tag{2.10}$$

Proof. As before we conclude that A, S and Y have full rank. Using $A^+AA^H = A^H$, $A^+ = Y^+S$, $A^+A = Y^+Y$, $Y^+ = Y^H(YY^H)^{-1}$ and $(YY^H)^{-1} = (I-E)^{-1}$ yields

$$\begin{aligned} A^+b - \widetilde{x} - \varrho_{\widetilde{w}} &= A^+\varrho_{\widetilde{x}} + A^+A\widetilde{x} - A^H\widetilde{w} \\ &= A^+\varrho_{\widetilde{x}} - A^+A\varrho_{\widetilde{w}} \\ &= Y^+S\varrho_{\widetilde{x}} - Y^+Y\varrho_{\widetilde{w}} \\ &= Y^H(YY^H)^{-1}\delta \\ &= Y^H(I-E)^{-1}\delta\ . \end{aligned} \tag{2.11}$$

As in the proof of Theorem 2.1 we use (1.7), Lemma 1.1, (1.8) and (2.6) to prove the result. □

Again the equality (2.8) is the main part of the theorem. It is formulated in a way such that the derived estimates (2.9) and (2.10) are of good quality. In case of underdetermined linear systems the quality can be expected to be often better than for least squares problems: $|S|\mathbf{e}$ could be replaced by $|Y^H|\mathbf{e}$, where $\mathrm{cond}(S) \approx \mathrm{cond}(A)$, but Y is nearly unitary.

The other remarks following Theorem 2.1 apply accordingly, where now $A^T \approx QR$ and S is an approximate inverse of R^T. In particular, as has been mentioned before, good bounds rely on good approximations.

3 Computational results

In the following we report computational results, all performed in IEEE 754 double precision arithmetic [8] equivalent to about 16 decimal digits precision in Matlab [9]. To obtain mathematically rigorous results, the estimates are bounded by interval arithmetic using INTLAB [13], the Matlab toolbox for reliable computing. For least squares problems and underdetermined linear systems we compare the bounds

$$\begin{array}{ll} \text{[Ru12]} & \text{Normwise bounds (1.2) and (1.4) } \big[\,\text{(4.8) and (3.8) in [14]}\,\big], \\ \text{[Mi12]} & \text{Componentwise bounds (1.3) and (1.5) } \big[\,\text{taken from [11]}\,\big], \\ \text{new1} & \text{Componentwise bounds as in (2.3) and (2.9)}, \\ \text{new2} & \text{Componentwise bounds as in (2.4) and (2.10)}. \end{array} \tag{3.1}$$

For the new bounds always the minimum is taken of the left and the right bound in (2.3), (2.9), (2.4) and (2.10), respectively. The new bounds (2.4) and (2.10) are implemented in the routine `verifylss` in Version 7 of INTLAB.

All bounds for all methods in (3.1) rely solely on $\widetilde{x}, \widetilde{w}$ and S. Based on those floating-point quantities, interval arithmetic is used to compute rigorous bounds for the other quantities $X, Y, E, \alpha, \rho_{\widetilde{x}}, \rho_{\widetilde{w}}$ and δ and for computing the final bounds. Thus the prerequisites for all methods are the same. In particular all methods succeed or fail to compute rigorous bounds depending on α being strictly less than one or not.

For all methods S is an approximate inverse (by Matlab's `inv`) of the factor R or R^T of the approximate QR-factorization of A or A^T, respectively, and $\widetilde{x}$ and $\widetilde{w}$ are improved by the residual iterations (5.7) and (5.3) in [14], respectively. As has been mentioned earlier we improve the quality of all bounds by using $\widetilde{x}_1+\widetilde{x}_2$ and $\widetilde{w}_1 + \widetilde{w}_2$ for least squares and for underdetermined problems, respectively. Then accurate dot products are used to compute inclusions of the residuals $\rho_{\widetilde{x}} = b - A\widetilde{x}_1 - A\widetilde{x}_2 + \widetilde{w}$ and $\rho_{\widetilde{w}} = A^H\widetilde{w}$ in case of least squares problems, and similarly for underdetermined systems. This is the common base for all methods in (3.1).

For an interval $[a, b] \neq 0$ we define the "number of correct digits" by $-\log_{10}[(b-a)/|a+b|]$. We say an interval with a, b being adjacent double precision floating-point numbers is of "maximum accuracy". For such intervals the number of correct digits is between 15.65 and 15.95, depending on the distance to the next power of 2.

We first test random least squares problems with full matrix of different dimensions and condition numbers. Random rectangular matrices of specified condition number are generated via singular values [7]. Since verified lower and upper bounds for the solution vector are calculated, we can display in Table 3.1 the minimum number and median number of correct digits of the componentwise inclusions. The median number of correct digits of the new methods is often close to maximum accuracy, so we refrain from displaying the maximum.

For 1000 test cases each we test the four methods. Then the minimum and median number of correct digits of all $1000m$ solution components is displayed. As can be seen, Miyajima's bounds are better than [Ru12] for well-conditioned problems, and worse for ill-conditioned problems. The new bounds are always at least as good as both the previous ones, often near maximum accuracy. As expected, the second new bound is never worse than the first one.

For underdetermined linear systems the results are displayed in Table 3.2. They are, as expected, in general better than for least squares problems. Now Miyajima's componentwise estimates are always better than

Table 3.1 Computational results for random least squares problems of the methods in (3.1).

			min # corr. digits				median # corr. digits			
m	n	cond(A)	[Ru12]	[Mi12]	new1	new2	[Ru12]	[Mi12]	new1	new2
1000	50	1e2	11.7	15.7	15.7	15.7	15.3	15.8	15.8	15.8
1000	50	1e5	10.4	15.4	15.7	15.7	15.3	15.8	15.8	15.8
1000	50	1e10	10.5	5.8	12.1	13.9	15.3	11.1	15.8	15.8
1000	50	1e11	10.3	4.2	12.1	12.7	15.3	9.2	15.8	15.8
1000	50	1e12	5.8	2.8	5.8	7.7	15.1	7.2	15.7	15.8
1000	50	1e13	0.0	0.0	0.0	0.1	8.6	5.1	8.9	10.4
1000	100	1e2	10.9	15.7	15.7	15.7	15.3	15.8	15.8	15.8
1000	100	1e5	10.2	14.8	15.7	15.7	15.3	15.8	15.8	15.8
1000	100	1e10	11.4	6.3	13.0	14.3	15.2	10.4	15.8	15.8
1000	100	1e11	10.5	3.8	12.1	12.9	15.2	8.5	15.8	15.8
1000	100	1e12	4.8	2.0	4.9	6.5	15.1	6.5	15.7	15.8
1000	100	1e13	0.0	0.0	0.0	0.0	6.9	4.4	7.2	8.4
1000	200	1e2	10.3	15.7	15.7	15.7	15.3	15.8	15.8	15.8
1000	200	1e5	10.6	14.2	15.7	15.7	15.3	15.8	15.8	15.8
1000	200	1e10	10.3	4.3	12.1	13.1	15.2	9.6	15.8	15.8
1000	200	1e11	10.2	2.4	11.4	12.0	15.2	7.7	15.8	15.8
1000	200	1e12	4.0	0.0	4.0	5.4	15.0	5.7	15.4	15.7
1000	200	1e13	0.0	0.0	0.0	0.0	4.3	3.3	4.7	5.6

Table 3.2 Computational results for random underdetermined linear systems of the methods in (3.1).

			min # corr. digits				median # corr. digits			
n	m	cond(A)	[Ru12]	[Mi12]	new1	new2	[Ru12]	[Mi12]	new1	new2
50	1000	1e2	7.2	8.5	15.7	15.7	12.7	14.0	15.8	15.8
50	1000	1e5	6.4	7.7	15.7	15.7	12.7	14.0	15.8	15.8
50	1000	1e10	6.9	8.2	12.2	13.1	12.6	14.0	15.8	15.8
50	1000	1e11	6.7	8.1	12.2	12.8	12.6	14.0	15.8	15.8
50	1000	1e12	6.2	7.7	10.9	11.6	12.6	14.0	15.8	15.8
50	1000	1e13	2.8	4.4	4.1	5.1	12.5	13.9	15.1	15.7
100	1000	1e2	7.3	8.5	15.7	15.7	12.8	14.0	15.8	15.8
100	1000	1e5	6.7	7.9	15.7	15.7	12.7	14.0	15.8	15.8
100	1000	1e10	6.4	7.9	12.8	13.1	12.7	14.0	15.8	15.8
100	1000	1e11	6.6	7.9	11.8	12.2	12.7	14.0	15.8	15.8
100	1000	1e12	7.6	8.9	11.5	12.1	12.7	14.0	15.8	15.8
100	1000	1e13	4.2	5.9	5.3	6.3	12.5	13.9	14.8	15.5
200	1000	1e2	7.1	8.2	15.7	15.7	12.9	14.0	15.8	15.8
200	1000	1e5	7.3	8.6	15.7	15.7	12.8	14.0	15.8	15.8
200	1000	1e10	5.9	7.3	11.8	12.1	12.7	14.0	15.8	15.8
200	1000	1e11	6.6	8.0	11.8	12.0	12.7	14.0	15.8	15.8
200	1000	1e12	6.9	8.1	10.4	10.7	12.7	14.0	15.7	15.8
200	1000	1e13	0.0	0.0	0.0	0.0	12.0	13.6	13.3	14.2

the previous bounds in [Ru12], and with three exceptions the first new bound is better than Miyajima's, whereas the second new bound is again never worse than the first.

Finally we display results for larger sparse problems. Note that R and also S are of size $n \times n$, however, $X = AS$ and $Y = SA$ for over- and underdetermined systems, respectively, are of size $m \times n$ and usually full. The direct computation of these full matrices can be avoided as described in [14, Section 6]. Sparse test matrices are taken from the Florida sparse matrix collection [3]. We use examples for least squares problems and for underdetermined linear systems for both tests by treating A and A^T, respectively. For the sparse problems we display only the minimum number of correct digits. As can be seen in Tables 3.3 and 3.4 the new methods compute in all examples inclusions of full accuracy. For least squares problems there is no difference to Miyajima's bounds as in (1.3), for underdetermined problems our bounds are better.

The computing time of all methods is essentially proportional to the time to compute the economy-size QR-decomposition, which requires $\mathcal{O}(mn^2)$ floating-point operations. Note this is significantly better than the $\mathcal{O}((m+n)^3)$ flops to solve (1.1), in particular if $m \gg n$.

Table 3.3 Computational results for sparse least squares problems of the methods in (3.1).

				min # corr. digits			
m	n	density[%]	Matrix	[Ru12]	[Mi12]	new1	new2
37932	331	1.09	JGD_Taha/abtaha2	13.4	15.7	15.7	15.7
14596	209	1.68	JGD_Taha/abtaha1	11.6	15.7	15.7	15.7
29493	11822	0.03	Sumner/graphics	10.5	15.2	15.7	15.7
10595	4929	0.09	HB/gemat1	7.8	15.7	15.7	15.7
12061	2262	0.09	LPnetlib/lp_80bau3b	9.2	15.7	15.7	15.7
13525	3000	0.12	LPnetlib/lp_fit2p	11.6	15.7	15.7	15.7
25067	1118	0.52	LPnetlib/lp_osa_07	11.7	15.7	15.7	15.7
54797	2337	0.25	LPnetlib/lp_osa_14	11.1	15.7	15.7	15.7
63516	507	1.27	Mittelmann/rail507	11.2	15.7	15.7	15.7
10757	124	6.82	Meszaros/air03	12.8	15.7	15.7	15.7
16819	4400	0.20	Meszaros/model10	8.6	15.7	15.7	15.7
123409	73	10.04	Meszaros/nw14	12.4	15.7	15.7	15.7
61521	4050	0.11	Meszaros/rlfddd	10.2	15.7	15.7	15.7
63076	3173	0.25	Meszaros/stat96v4	8.7	15.7	15.7	15.7
184756	190	23.68	JGD_BIBD/bibd_20_10	12.6	15.7	15.7	15.7
319770	231	12.12	JGD_BIBD/bibd_22_8	12.0	15.7	15.7	15.7

Table 3.4 Computational results for sparse underdetermined linear systems of the methods in (3.1).

				min # corr. digits			
n	m	density[%]	Matrix	[Ru12]	[Mi12]	new1	new2
331	37932	1.09	JGD_Taha/abtaha2	8.9	9.8	15.7	15.7
209	14596	1.68	JGD_Taha/abtaha1	7.3	8.5	15.7	15.7
11822	29493	0.03	Sumner/graphics	6.9	8.1	15.7	15.7
4929	10595	0.09	HB/gemat1	6.5	8.2	15.7	15.7
2262	12061	0.09	LPnetlib/lp_80bau3b	9.5	10.5	15.7	15.7
3000	13525	0.12	LPnetlib/lp_fit2p	9.8	10.5	15.7	15.7
1118	25067	0.52	LPnetlib/lp_osa_07	7.5	9.2	15.7	15.7
2337	54797	0.25	LPnetlib/lp_osa_14	7.4	9.0	15.7	15.7
507	63516	1.27	Mittelmann/rail507	6.4	8.2	15.7	15.7
124	10757	6.82	Meszaros/air03	6.1	7.6	15.7	15.7
4400	16819	0.20	Meszaros/model10	7.0	8.4	15.7	15.7
73	123409	10.04	Meszaros/nw14	3.8	6.1	15.7	15.7
4050	61521	0.11	Meszaros/rlfddd	5.1	6.4	15.7	15.7
3173	63076	0.25	Meszaros/stat96v4	7.5	8.2	15.7	15.7
190	184756	23.68	JGD_BIBD/bibd_20_10	7.6	8.2	15.7	15.7
231	319770	12.12	JGD_BIBD/bibd_22_8	6.9	7.7	15.7	15.7

4 Appendix

We try to explain a phenomenon which has been noted in [14, Section 7, in particular Figure 7.1]. Sometimes, as in [1,4], the augmented linear system (1.1) is symmetrized into

$$\begin{pmatrix} -I & A \\ A^H & \mathbf{0} \end{pmatrix} \begin{pmatrix} w \\ x \end{pmatrix} = \begin{pmatrix} b \\ \mathbf{0} \end{pmatrix} , \tag{4.1}$$

Suppose the condition number of a matrix C is 10^k, and an approximate solution $\widetilde{x}$ of a linear system $Cx = c$ is obtained in double precision by Gaussian elimination with partial pivoting. Then, according to the well-known rule of thumb in numerical analysis, the number of correct digits of $\widetilde{x}$ should be about $16 - k$. Practical evidence suggests that this is indeed true for the augmented system (4.1).

The matrix of the augmented system (1.1) exchanges the two block columns of the matrix, so the condition number does not change. However, practical experience suggests that the solution of (1.1) has significantly more than $16 - k$ correct digits, contradicting the mentioned rule of thumb.

As a typical example, we generate a random 500 × 100 matrix $\mathtt{A}$ with condition number $\mathrm{cond}_2(A) = \sigma_{\max}(A)/\sigma_{\min}(A) = 10^{10}$ with right hand side $\mathtt{b} = \mathbf{e}$. The true condition number of both the matrix in

Table 4.1 Sample code to test the augmented linear systems (1.1) and (4.1).

```
X = verifylss(A,b);         % verified inclusion of the true solution
disp(' ')
accX = max(relerr(X))
disp(' ')

disp('symmetric augmented matrix')
B = [-eye(m) A;A' zeros(n)];
xsym = B\[b;zeros(n,1)];
x = xsym(m+1:end);
relerrsym = min(relerr(x,X))
disp(' ')

disp('unsymmetric augmented matrix')
B = [A -eye(m);zeros(n) A'];
xunsym = B\[b;zeros(n,1)];
x = xunsym(1:n);
relerrunsym = median(relerr(x,X))
```

Table 4.2 Sample result of the code in Table 4.1 testing the augmented linear systems (1.1) and (4.1).

```
accX =
  2.2076e-016

symmetric augmented matrix
relerrsym =
     1

unsymmetric augmented matrix
relerrunsym =
  1.5538e-004
```

the symmetric (4.1) and the unsymmetric system (1.1) is 10^{20}, computed by some multiple-precision package, so that an approximate solution is expected to have no correct digit.

Algorithm `verifylss` in INTLAB implements our new methods and computes an inclusion `X` of the true solution (see the code in Table 4.1). Note that this includes the proof that `A` has full rank. As can be seen in the displayed results, the maximum relative error $2 \cdot 10^{-16}$ of the inclusion `X` implies that all components of the inclusion `X` are correct to the last digit. Next (cf. Table 4.1) the symmetric linear system (4.1) is generated and solved, producing the approximate solution `xsym`. The median relative error `relerrsym` (see Table 4.2) of the relevant components against the inclusion is 1, which means that, as expected, in the median the approximate solution of the symmetric system (4.1) has no correct digit.

Finally, the unsymmetric linear system (1.1) is generated and solved, producing the approximate solution `xunsym`. The median relative error `relerrunsym` of the relevant components against the inclusion is about 10^{-4}, which means that in the median the approximate solution of the unsymmetric system has about 4 correct digits. This contradicts the mentioned rule of thumb.

The reason seems to be the following. Suppose that A is equilibrated with a norm of order 1. The backslash operator in Matlab uses Gaussian elimination with partial pivoting. Thus in the first elimination block in (1.1) only pivots of the matrix A are used, whereas in the symmetric system (4.1) pivots out of the identity are used. Seemingly this destroys the structure, so that the ill-conditioning of the symmetric system appears.

Table 4.3 Sample Matlab/INTLAB code to compute the bounds (1.2), (1.3), (2.3) and (2.4) for least squares problems.

```
function D = test_lsqr(A,b)
% output number of correct digits for [Ru12], [Mi12], new1 (2.3) and new2 (2.4)
  n = min(size(A));
  % Generation of input data. All bounds are based on the same data.
  R = qr(A,0);                      % economy-size qr-decomposition
  R = triu(R(1:n,:));               % choose upper part
  S = inv(full(R));                 % approximate inverse of R
  X = A*intval(S);                  % inclusion of AS
  E = eye(n)-X'*X;                  % inclusion of I-X'*X
  alpha = norm(E,inf);              % bound for ||I-X'X||_inf
  if ~(alpha<1)                     % no inclusion possible
    D = NaN(n,4);
    return
  end
  [xs1,xs2,ws] = resid_iter_lsqr(A,b,S,mid(X));  % residual correction to produce xs and ws
  res_ws = Dot_(A',ws,-2);                  % inclusion of residuals
  res_xs = Dot_(1,b,A,-xs1,A,-xs2,1,ws,-2);
  % first bound [Ru12]
  t1 = norm(S*(X'*res_xs),inf);             % inclusions of terms
  t2 = norm(S*(S'*res_ws),inf);
  t31 = alpha*norm(intval(S),inf)/(1-alpha);
  t32 = norm(X'*res_xs,inf) + norm(S'*res_ws,inf);
  BoundRu12 = xs1 + ( intval(xs2) + midrad( 0 , mag( t1 + t2 + t31*t32 ) ) );
  % second bound [Mi12]
  t1 = abs(S*(S'*(A'*res_xs-res_ws)));      % inclusions of terms
  t21 = norm(S'*(A'*res_xs-res_ws),inf)/(1-alpha);
  t22 = abs(S) * ( abs(E)*ones(n,1) );
  BoundMi12 = xs1 + ( intval(xs2) + midrad( 0 , mag( t1 + t21.*t22 ) ) );
  % new bound (2.3)
  delta = X'*res_xs - S'*res_ws;            % inclusions of terms
  b1 = mag( norm(delta,inf)/(1-alpha) * ( abs(S)*intval(ones(n,1)) ) );
  b2 = mag( norm(delta,2)/(1-alpha) * sqrt(sum(S.*intval(S),2)) );
  Boundnew1 = xs1 + ( intval(xs2) + midrad(0,min(b1,b2)) );
  % new bound (2.4)
  delta = X'*res_xs - S'*res_ws;            % inclusions of terms
  b1 = mag( norm(E*delta,inf)/(1-alpha) * ( abs(S)*intval(ones(n,1)) ) );
  b2 = mag( norm(E*delta,2)/(1-alpha) * sqrt(sum(S.*intval(S),2)) );
  Boundnew2 = xs1 + ( intval(xs2) + ( S*delta + midrad(0,min(b1,b2)) ) );
  % number of correct digits of bounds
  D = [BoundRu12 BoundMi12 Boundnew1 Boundnew2];
  D = -log10(relerr(D));
```

Another statement, which is generally true, also proves to have exceptions. Usually total pivoting produces more accurate results than partial pivoting. More precisely, vast practical experience shows that Gaussian elimination with partial pivoting is in most cases a stable algorithm, but theoretically it is highly unstable due to the possibility of exponential pivot growth.[3]

In our example, Gaussian elimination with total pivoting produces for both the unsymmetric system (1.1) and the symmetric system (4.1) the same result because the matrices are permutations of each other. However, it is likely that the result with total pivoting is the same as for partial pivoting of the symmetric system.

[3] As is known, there are instances where this happens in practice as well [19].

Thus the approximation produced by total pivoting has no correct digit, in contrast to partial pivoting with about 4 correct digits.

These statements seem to be true in general, not just for the displayed example. Note that the condition number of the least squares problem [18] is basically $\mathrm{cond}(A)^2$ times the norm of the residual $A\widetilde{x}-b$; the point here was to show that the general rule of thumb for linear systems does not necessarily apply to structured matrices, so that using the unsymmetric system (1.1) is preferable to the symmetrized system (4.1).

Table 4.3 shows the sample Matlab/INTLAB code to compute the bounds (1.2), (1.3), (2.3) and (2.4) for least squares problems. The routine `resid_iter_lsqr` implements the residual iteration (5.7) in [14], `Dot_` is an INTLAB routine to compute accurate approximations or inclusions of dot product expressions.

Acknowledgment

The author thanks Florian Bünger for his fruitful comments on a preliminary version of this manuscript, and two anonymous referees for their thorough reading and stimulating remarks.

References

1. M. Arioli, I.S. Duff, and P.P.M. de Rijk. On the augmented system approach to sparse least-squares problems. *Numerische Mathematik*, 55(6):667–687, 1989.
2. Å. Björck. Iterative refinement of linear least squares solutions I. *BIT*, 7:257–278, 1967.
3. T.A. Davis and Y. Hu. The University of Florida Sparse Matrix Collection. *ACM Transactions on Mathematical Software*, 38(1):1:1–1:25, 2011.
4. J.B. Demmel, Y. Hida, W. Kahan, X.S. Li, S. Mukherjee, and E.J. Riedy. Error Bounds from Extra Precise Iterative Refinement. *ACM Transactions on Mathematical Software (TOMS)*, 32(2):325–351, 2006.
5. A. Frommer. Proving Conjectures by Use of Interval Arithmetic. In U. Kulisch et al., editor, *Perspectives on enclosure methods. SCAN 2000, GAMM-IMACS international symposium on scientific computing, computer arithmetic and validated numerics, Univ. Karlsruhe, Germany, September 19-22, 2000*, Wien, 2001. Springer.
6. G.H. Golub and Ch. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, third edition, 1996.
7. N. J. Higham. *Accuracy and stability of numerical algorithms.* SIAM Publications, Philadelphia, 2nd edition, 2002.
8. *ANSI/IEEE 754-2008: IEEE Standard for Floating-Point Arithmetic.* New York, 2008.
9. MATLAB. User's Guide, Version 7, the MathWorks Inc., 2004.
10. S. Miyajima. Fast enclosure for solutions in underdetermined systems. *Journal of Computational and Applied Mathematics (JCAM)*, 234:3436–3444, 2010.
11. S. Miyajima. Componentwise enclosure for solutions in least squares problems and underdetermined linear systems. SCAN conference Novosibirsk, 2012.
12. S.M. Rump. *Kleine Fehlerschranken bei Matrixproblemen.* PhD thesis, Universität Karlsruhe, 1980.
13. S.M. Rump. INTLAB - INTerval LABoratory. In Tibor Csendes, editor, *Developments in Reliable Computing*, pages 77–104. Kluwer Academic Publishers, Dordrecht, 1999.
14. S.M. Rump. Verified Bounds for Least Squares Problems and Underdetermined Linear Systems. *SIAM J. Matrix Anal. Appl. (SIMAX)*, 33(1):130–148, 2012.
15. N.V. Sahinidis and M. Tawaralani. A polyhedral branch-and-cut approach to global optimization. *Math. Programming*, B103:225–249, 2005.
16. H.J. Stetter. Sequential Defect Correction in High-Accuracy Floating-Point Arithmetics. *Numerical Analysis*, 1066:186–202, 1984. Proceedings, Dundee 1983.
17. W. Tucker. The Lorenz attractor exists. *C. R. Acad. Sci., Paris, Sér. I, Math.*, 328(12):1197–1202, 1999.
18. P.Å. Wedin. Perturbation theory for pseudo-inverses. *BIT*, 13:217–232, 1973.
19. S.J. Wright. A collection of problems for which Gaussian elimination with partial pivoting is unstable. *SIAM J. Sci. Comput. (SISC)*, 14(1):231–238, 1993.