NOLTA, IEICE

Paper

Computable backward error bounds for basic algorithms in linear algebra

Siegfried M. Rump [1 a)]

[1] *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*

a) *rump@tuhh.de*

Received October 20, 2014; Revised January 9, 2015; Published July 1, 2015

Abstract: Standard error estimates in numerical linear algebra are often of the form $\gamma_k|R||S|$ where R, S are known matrices and $\gamma_k := k\mathbf{u}/(1-\mathbf{u})$ with $\mathbf{u}$ denoting the relative rounding error unit. Recently we showed that for a number of standard problems γ_k can be replaced by $k\mathbf{u}$ for any order of computation and without restriction on the dimension. Such problems include LU- and Cholesky decomposition, triangular system solving by substitution, matrix multiplication and more. The theoretical bound implies a practically computable bound by estimating the error in the floating-point computation of $k\mathbf{u}|R||S|$. Standard techniques, however, imply again a restriction on the dimension. In this note we derive simple computable bounds being valid without restriction on the dimension. As the bounds are mathematically rigorous, they may serve in computer assisted proofs.

Key Words: error estimates, numerical algorithms, LU factorization, Cholesky factorization, triangular system solving, summation, dot products, rounding error analysis, unit in the first place, computer assisted proofs

1. Introduction and main result

Denote by $\mathbb{F}$ a set of floating-point numbers with relative rounding error unit $\mathbf{u}$ and with operations according to IEEE 754 [1]. Throughout the paper we assume that no overflow nor underflow occurs. Then for an operation $\circ \in \{+, -, \cdot, /\}$ and $a, b \in \mathbb{F}$ the floating-point result $\mathrm{fl}(a \circ b)$ satisfies [2]

$$|\mathrm{fl}(a \circ b) - a \circ b| \leqslant \mathbf{u}|a \circ b|. \tag{1}$$

For matrices $A \in \mathbb{F}^{m\times k}$ and $B \in \mathbb{F}^{k\times n}$ denote by $\widehat{P}$ the floating-point result of the true product $P := AB$. Then [3]

$$|\widehat{P} - P| \leqslant k\mathbf{u}|A||B| \tag{2}$$

is true for any dimension $k \in \mathbb{N}$ and no matter what the order of evaluation. Thus the bound is valid for library routines including blocked code [4]; it does not apply to Strassen-like methods [5]. The bound improves the well-known factor $\gamma_k := k\mathbf{u}/(1 - k\mathbf{u})$ (cf. for example [2]) into $k\mathbf{u}$ and removes the implicit restriction on k. Similarly [6]

Nonlinear Theory and Its Applications, IEICE, vol. 6, no. 3, pp. 1–4 ©IEICE 2015 DOI: 10.1588/nolta.6.1

$$|A-\widehat{L}\widehat{U}| \leqslant n\mathbf{u}|\widehat{L}||\widehat{U}| \quad \text{and} \quad |A-\widehat{G}\widehat{G}^T| \leqslant (n+1)\mathbf{u}|\widehat{G}||\widehat{G}^T|$$

are valid for computed LU-factors $\widehat{L}, \widehat{U}$ or Cholesky factor $\widehat{G}$ of an $n \times n$ matrix A, respectively. Denoting the floating-point product of $|A||B|$ by $\widehat{Q}$, (2) can be used to estimate the error in the computation of $|A||B|$. The standard approach is

$$|A||B| =: Q \leqslant \widehat{Q} + |\widehat{Q}-Q| \leqslant \widehat{Q} + k\mathbf{u}Q$$

and, provided $k\mathbf{u} < 1$,

$$|\mathrm{fl}(AB)-AB| \leqslant \frac{k\mathbf{u}}{1-k\mathbf{u}}\mathrm{fl}(|A||B|). \tag{3}$$

The same applies to the other mentioned estimates. Moreover, it is not difficult to take care of underflow.

In the following, we remove the restriction on k, and we improve (3) by using the *unit in the first place* (ufp): a real number t being given, we have $\mathrm{ufp}(0) = 0$ and, if $t \neq 0$, $\mathrm{ufp}(t) := 2^{\lfloor \log_2 |t| \rfloor}$. Thus $\mathrm{ufp}(t)$ can be thought of as the weight of its first nonzero bit in its binary representation. Then (1) can be replaced by

$$|\mathrm{fl}(a \circ b) - a \circ b| \leqslant \mathbf{u}\mathrm{ufp}(a \circ b) \leqslant \mathbf{u}\mathrm{ufp}(\mathrm{fl}(a \circ b)). \tag{4}$$

Note that this improves on (1) by up to a factor of 2 depending on how close $|a \circ b|$ is to $\mathrm{ufp}(a \circ b)$.

The following result was proved in [7] for recursive summation. We now extend it to summation in any order.

Lemma 1 Let $x \in \mathbb{F}^k$ be given, define $s := \sum_{i=1}^k x_i$, denote by $\widehat{s}$ the floating-point sum of the x_i in any order, and denote by $\widehat{S}$ the floating-point sum of the absolute values $|x_i|$ computed in the same order as $\widehat{s}$. Then

$$|\widehat{s} - s| \leqslant (k-1)\mathbf{u} \cdot \mathrm{ufp}(\widehat{S}) \tag{5}$$

is true without restriction on k. If $48\mathbf{u} < 1$, then the estimate may be false if the ordering in the computation of $\widehat{s}$ and $\widehat{S}$ is not the same.

Proof. For the proof we proceed by induction. For $k = 1$ there is nothing to prove. Assume (5) is true for floating-point summation of up to $k-1$ summands in any order. Then $\widehat{s} = \mathrm{fl}(\widehat{s}_1 + \widehat{s}_2)$ for some disjoint splitting I_1, I_2 of the index set $\{1, \ldots, k\}$ with $k_\nu := |I_\nu|$, where $\widehat{s}_\nu$ is the floating-point sum of the x_i with $i \in I_\nu$ in some order for $\nu \in \{1, 2\}$. Define $s_\nu := \sum_{i \in I_\nu} s_i$, so that $s = s_1 + s_2$. By assumption, the floating-point summation of x_i and $|x_i|$ is performed in the same order, implying $|\widehat{s}| \leqslant \widehat{S}$ and therefore $\mathrm{ufp}(\widehat{s}) \leqslant \mathrm{ufp}(\widehat{S})$. Thus (4), the induction hypothesis, $k_1 + k_2 = k$ and $0 \leqslant \widehat{S}_1, \widehat{S}_2 \leqslant \widehat{S}$ yield

$$\begin{aligned} |\widehat{s} - s| &= |\widehat{s} - (\widehat{s}_1 + \widehat{s}_2) + \widehat{s}_1 - s_1 + \widehat{s}_2 - s_2| \\ &\leqslant \mathbf{u}\mathrm{ufp}(\widehat{s}) + (k_1 - 1)\mathbf{u}\mathrm{ufp}(\widehat{S}_1) + (k_2 - 1)\mathbf{u}\mathrm{ufp}(\widehat{S}_2) \\ &\leqslant \mathbf{u}\mathrm{ufp}(\widehat{S}) + (k-2)\mathbf{u}\mathrm{ufp}(\widehat{S}), \end{aligned}$$

and thus (5). It is mandatory that $\widehat{s}$ and $\widehat{S}$ are computed in the same order as by the following example:

$$x_1 = 1 - 4\mathbf{u},\ x_{2\ldots5} = \mathbf{u}/2(1+8\mathbf{u}),\ x_6 = \mathbf{u}(1+8\mathbf{u}).$$

Note that all x_i are in $\mathbb{F}$. Denoting by $\widehat{s}_\nu$ the partial sums of recursive summation we have $\widehat{s}_2 = 1 - 3\mathbf{u}$, $\widehat{s}_3 = 1 - 2\mathbf{u}$, $\widehat{s}_4 = 1 - \mathbf{u}$, $\widehat{s}_5 = 1$ and $\widehat{s} = \widehat{s}_6 = 1 + 2\mathbf{u}$, so that

$$|\widehat{s} - s| = 1 + 2\mathbf{u} - [1 - 4\mathbf{u} + 4\mathbf{u}/2(1+8\mathbf{u}) + \mathbf{u}(1+8\mathbf{u})] = 3\mathbf{u} - 24\mathbf{u}^2.$$

But summing $x_{6\ldots2}$ for the error estimate yields $\widehat{t} := 3\mathbf{u} + 24\mathbf{u}^2$ without rounding error and $\widehat{S} = \mathrm{fl}(x_1 + \widehat{t}) = 1 - \mathbf{u}$ with $\mathrm{ufp}(\widehat{S}) = 0.5$. Thus, using $48\mathbf{u} < 1$,

$$|\widehat{s} - s| = 3\mathbf{u} - 24\mathbf{u}^2 > 5\mathbf{u}\mathrm{ufp}(\widehat{S})$$

contradicting (5). □

Note that (5) is a computable bound because $\mathrm{ufp}(\widehat{S})$ can be computed with Algorithm 3.6 in [7] with four floating-point operations in rounding to nearest.

However, an obstacle is that the original sum and the sum of absolute values for the error bound have to be computed in the same order. We remove that assumption by the following simple and computable estimate for summation and dot products.

Lemma 2 Let non-negative real $0 \leqslant x \in \mathbb{R}^k$ be given with $\widehat{x}_i := \mathrm{fl}(x_i)$ denoting the rounding of x_i into $\mathbb{F}$. Denote by $S := \sum_{i=1}^k x_i$ the true sum of the real x_i, and by $\widehat{S}$ the floating-point sum of the $\widehat{x}_i$ in any order. Then, without restriction on $k \in \mathbb{N}$,

$$S \leqslant (1+\mathbf{u})(\widehat{S} + (k-1)\mathbf{u}\mathrm{ufp}(\widehat{S})). \tag{6}$$

At least for recursive summation and any vector length k equality may be attained. If $x \in \mathbb{F}^k$, then

$$S \leqslant \widehat{S} + (k-1)\mathbf{u}\mathrm{ufp}(\widehat{S}). \tag{7}$$

is true without restriction on k, and equality can be attained for any vector length k.

Remark. We note that for special summation schemes, in particular binary summation, sharper estimates can be derived.

Proof. Assertion (7) follows by Lemma 1. Equality is attained for $x_1 := 1$ and $x_{2\ldots k} := \mathbf{u}$ for $k \geqslant 1$, for which $\widehat{S} = 1$ for every $k \geqslant 1$.

If $0 \leqslant x_i \in \mathbb{R}$, then Lemma 1, (4) and (7) yield

$$\begin{aligned} S &= \widehat{S} + \textstyle\sum_{i=1}^k \widehat{x}_i - \widehat{S} + \sum_{i=1}^k (x_i - \widehat{x}_i) \\ &\leqslant \widehat{S} + (k-1)\mathbf{u}\mathrm{ufp}(\widehat{S}) + \mathbf{u}\textstyle\sum_{i=1}^k \widehat{x}_i \\ &\leqslant \widehat{S} + (k-1)\mathbf{u}\mathrm{ufp}(\widehat{S}) + \mathbf{u}\left(\widehat{S} + (k-1)\mathbf{u}\mathrm{ufp}(\widehat{S})\right) \end{aligned}$$

and prove (6). Equality is attained for $x_1 := 1+\mathbf{u}$ and $x_{2\ldots k} := \mathbf{u}(1+\mathbf{u})$ for $k \geqslant 1$. Then $\widehat{x}_1 = 1$ and $\widehat{x}_i = \mathbf{u}$ for all $2 \leqslant i \leqslant k$, so that again $\widehat{S} = 1$ for every $k \geqslant 1$. □

Corollary 1 Let $A \in \mathbb{F}^{m\times k}$ and $B \in \mathbb{F}^{k\times n}$ be given, denote the floating-point computation of AB by $\widehat{P}$, and denote the floating-point computation of $|A||B|$ by $\widehat{Q}$. Both $\widehat{P}$ and $\widehat{Q}$ may be computed in any, not necessarily the same order. Then, without restriction on $k \in \mathbb{N}$,

$$|\widehat{P} - AB| \leqslant k\mathbf{u}(1+\mathbf{u})(\widehat{Q} + (k-1)\mathbf{u}\mathrm{ufp}(\widehat{Q})). \tag{8}$$

Proof. Combining (2) and Lemma 2 proves the result. □

Recall that $\mathrm{ufp}(\widehat{Q})$ is computed by Algorithm 3.6 in [7] solely in floating-point arithmetic, and that it is straightforward to include possible underflow in the estimate. Moreover, inevitable rounding errors in the floating-point computation of an upper bound of the right hand side in (6) or (7) are easily estimated using (1) or (4).

We finally mention a computable lower bound on s depending only on $\widehat{s}$ and k. It might be less useful in practice, however, it shows the quality of the bound in (7).

Lemma 3 Let $0 \leqslant x \in \mathbb{F}^k$ be given. Denote by $S := \sum_{i=1}^k x_i$ the true sum, and by $\widehat{S}$ the floating-point sum in any order. Then, without restriction on $k \in \mathbb{N}$,

$$\frac{\widehat{S}}{(1+\mathbf{v})^{k-1}} \leqslant S, \tag{9}$$

where $\mathbf{v} := \mathbf{u}/(1+\mathbf{u})$. Note that $(1+\mathbf{v})^{-k+1} \geqslant 1-(k-1)\mathbf{v}$.

Proof. For $t \in \mathbb{R}$ the sharper estimate

$$|\mathrm{fl}(t) - t| \leqslant \mathbf{v}|t| \tag{10}$$

was noted in [5], improving upon (1). This implies

$$\mathrm{fl}(a+b) \leqslant (1+\mathbf{v})(a+b) \qquad \text{and} \qquad a+b \leqslant (1+\mathbf{u})\mathrm{fl}(a+b). \tag{11}$$

for $0 \leqslant a, b \in \mathbb{F}$.

For $k = 1$ there is nothing to prove. Assume (9) is true for the floating-point summation of up to $k-1$ summands in any order. As in the proof of Lemma 2 let $\widehat{S} = \mathrm{fl}(\widehat{S}_1 + \widehat{S}_2)$ for some disjoint splitting I_1, I_2 of the index set $\{1, \ldots, k\}$ with $k_\nu := |I_\nu|$ and $\widehat{S}_\nu$ denoting the floating-point summation of the x_i with $i \in I_\nu$ in some order, and S_ν denoting the true sum. Then the induction hypothesis, $k_1 + k_2 = k$, $\mathbf{v} \geqslant 0$, $k_1, k_2 \geqslant 1$ and (11) yield

$$\begin{aligned} S &= S_1 + S_2 \\ &\geqslant (1+\mathbf{v})^{1-k_1}\widehat{S}_1 + (1+\mathbf{v})^{1-k_2}\widehat{S}_2 \\ &= (1+\mathbf{v})^{1-k}\left[(1+\mathbf{v})^{k_2}\,\widehat{S}_1 + (1+\mathbf{v})^{k_1}\,\widehat{S}_2\right] \\ &\geqslant (1+\mathbf{v})^{1-k} \cdot (1+\mathbf{v})(\widehat{S}_1 + \widehat{S}_2) \\ &\geqslant (1+\mathbf{v})^{1-k} \cdot \widehat{S} \end{aligned}$$

□

Acknowledgments

The author wishes to thank the two anonymous referees for their constructive and valuable comments.

References

[1] *ANSI/IEEE 754-2008: IEEE Standard for Floating-Point Arithmetic*, IEEE, New York, 2008.

[2] N.J. Higham, *Accuracy and stability of numerical algorithms*, SIAM Publications, Philadelphia, 2nd edition, 2002.

[3] C.-P. Jeannerod and S.M. Rump, "Improved error bounds for inner products in floating-point artihmetic," *SIAM. J. Matrix Anal. & Appl. (SIMAX)*, vol. 34, no. 2, pp. 338–344, 2013.

[4] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, and S. Hammarling, *LAPACK User's Guide*, SIAM Publications, Philadelphia, 1992.

[5] D.E. Knuth, *The Art of Computer Programming: Seminumerical Algorithms*, vol. 2, Addison Wesley, Reading, Massachusetts, second edition, 1981.

[6] S.M. Rump and C.-P. Jeannerod, "Improved backward error bounds for LU and Cholesky factorizations," *SIAM. J. Matrix Anal. & Appl. (SIMAX)*, vol. 35, no. 2, pp. 684–698, 2014.

[7] S.M. Rump, "Error estimation of floating-point summation and dot product," *BIT Numerical Mathematics*, vol. 52, no. 1, pp. 201–220, 2012.