Numer. Math. 27, 485–490 (1977)

Backward Error Analysis for Totally Positive Linear Systems*

Carl de Boor and Allan Pinkus

Received September 24, 1976

Summary. Gauss elimination applied to an $n \times n$ matrix A in floating point arithmetic produces (if successful) a factorization $\hat{L}\hat{U}$ which differs from A by no more than $\gamma|\hat{L}|\,|\hat{U}|$, for some γ of order n times the unit roundoff. If A is totally positive, then both computed factors $\hat{L}$ and $\hat{U}$ are nonnegative for sufficiently small unit roundoff and one obtains pleasantly small bounds for the perturbation in A which would account for the rounding errors committed in solving $Ax=b$ for x by Gauss elimination *without pivoting*. It follows that the banded linear system for the B-spline coefficients of an interpolating spline function can be solved safely by Gauss elimination without pivoting.

1. Introduction

It is possible to state the result of Wilkinson's backward error analysis of Gauss elimination without pivoting as follows: The triangular factors $\hat{L}$ and $\hat{U}$ for the given $n \times n$ matrix A, as computed in some $t\,\beta$-digit floating point arithmetic, satisfy

$$\hat{L}\hat{U} = A + E \quad \text{with } |E| \leqq \gamma |\hat{L}|\,|\hat{U}|. \tag{1}$$

Here, $\gamma = n\,u/(1-n\,u)$, and u is the unit roundoff. Further, the solution $\hat{x}$ for $Ax=b$, computed with the aid of the factorization $\hat{L}\hat{U} \cong A$, is the exact solution of

$$(\hat{L}\hat{U} + \tilde{E})\,\hat{x} = b \quad \text{with } |\tilde{E}| \leqq \gamma(1+\gamma)\,|\hat{L}|\,|\hat{U}|, \tag{2}$$

hence of

$$(A + \Delta A)\,\hat{x} = b \quad \text{with } |\Delta A| \leqq |E| + |\tilde{E}| \leqq \gamma(2+\gamma)\,|\hat{L}|\,|\hat{U}|. \tag{3}$$

These bounds are meant to be *pointwise*, i.e.,

$$B \leqq C \quad \text{iff for all } i, j,\ b_{ij} \leqq c_{ij},$$

and $|B|$ stands for the matrix of absolute values of the entries of B,

$$|B| := (|b_{ij}|).$$

To be precise, we use the model of $t\,\beta$-digit floating point arithmetic described in Forsythe and Moler [5] or in Stoer [8]: We assume that, with ω any of the four arithmetic operations $+, -, *, /$, the corresponding floating point operation $\hat{\omega}$ on the $t\,\beta$-digit floating point numbers $\mathbb{R}_{t,\beta}$ satisfies

$$x\,\hat{\omega}\,y = \mathrm{fl}(x\,\omega\,y) \quad \text{for all } x, y \in \mathbb{R}_{t,\beta}$$

with

$$\mathrm{fl}: \mathbb{R} \to \mathbb{R}_{t,\beta}$$

* Sponsored by the United States Army under Contract No. DAAG29-75-C-0024 and the National Science Foundation under Grant No. MPS72-00381 A01.

a given map on the reals (e.g., rounding or chopping). We will ignore the possibility of overflow or underflow and will think of the *unit roundoff*

$$u := \sup_{x\in\mathbb{R}} \{|\mathrm{fl}\, x - x|/|x|,\ |\mathrm{fl}\, x - x|/|\mathrm{fl}\, x|\}$$

as a number of the order of β^{-t}.

These assumptions allow the conclusion that, for all $x, y \in \mathbb{R}_{t,\beta}$, there exist $\delta, \eta \in [1-u, 1+u]$ such that

$$x\, \hat{\omega}\, y = (x\,\omega\, y)\ \delta = (x\,\omega\, y)/\eta,$$

and, from this fact, one easily derives estimates like (1) and (2); see, e.g., Sections 4.5 and 4.6 of Stoer [8]. Such arguments do show that we can choose $\gamma = k\,u/(1-k\,u)$ in case the matrix A is $(2k-1)$ — banded.

We note the fact (provable by induction on n) that if A is invertible and has a triangular factorization LU, then we can indeed compute $\hat{L}$ and $\hat{U}$ by Gauss elimination without pivoting for sufficiently small unit roundoff u, and

$$\hat{L} \to L, \qquad \hat{U} \to U \qquad \text{as } u \to 0. \tag{4}$$

The formulation (1)–(3) leads to pleasantly small bounds on the perturbation matrices E and $\tilde{E}$ in case A is totally positive since then both $\hat{L}$ and $\hat{U}$ are nonnegative (for sufficiently small unit roundoff), and therefore $|\hat{L}|\,|\hat{U}| = \hat{L}\,\hat{U} = A + E$.

It follows that the banded linear system for the B-spline coefficients of an interpolating spline function can be solved safely by Gauss elimination without pivoting.

2. On Estimating $|\hat{L}|\,|\hat{U}|$

In the customary backward error analysis, the size of the two factors $|\hat{L}|$ and $|\hat{U}|$ is considered separately. Partial or total pivoting is used to insure that the entries of $|\hat{L}|$ are all in $[0, 1]$, so that $\|\hat{L}\|_\infty \leq n$. Then one is left to worry about just how big the entries of $|\hat{U}|$ might be. Forsythe and Moler [5] follow Wilkinson and introduce the growth factor ϱ in terms of which

$$|\hat{u}_{ij}| \leq \varrho \|A\|_\infty. \tag{5}$$

Therefore

$$\|\,|\hat{L}|\ |\hat{U}|\,\|_\infty \leq \|\hat{L}\|_\infty \|\hat{U}\|_\infty \leq n^2 \varrho \|A\|_\infty. \tag{6}$$

Forsythe and Moler [5; Theorem 21.41] obtain from such considerations that $\hat{x}$ is the exact solution of the linear system

$$(A + \Delta A)\,\hat{x} = b \qquad \text{with } \|\Delta A\|_\infty \leq \gamma\,(n^2 + 3n)\,\varrho\,\|A\|_\infty. \tag{7}$$

To be precise, a combination of (1) and (6) yields only

$$\|E\|_\infty \leq \gamma\, n^2 \varrho \|A\|_\infty,$$

while Forsythe and Moler [5] in fact prove

$$\|E\|_\infty \leq n^2 \varrho\, u \|A\|_\infty.$$

This is connected with the fact that (5) may be a poor bound since their ϱ necessarily measures the relative size of *all* entries encountered during the various steps of Gauss elimination, not just the final entries. But their estimate of the error in backsolving *is* based on (6) and so makes that error the major contribution to the bound (7).

It is the underlying thesis of this note that it might be more profitable at times to consider the product

$$\hat{A} := |\hat{L}|\,|\hat{U}|$$

directly. (We are indebted to a referee for the observation that Erisman and Reid [4] also take this point of view.)

Consider, for example, the case when A is symmetric and positive definite. The squareroot free form of the Cholesky decomposition for A then produces a computed factorization for A which satisfies

$$\hat{U}^T \hat{D} \hat{U} = A + E \qquad \text{with } |E| \leqq \gamma |\hat{U}^T|\,\hat{D}\,|\hat{U}| \tag{1'}$$

with $\hat{U}$ upper triangular and $\hat{D} := (\operatorname{diag} \hat{U})^{-1} > 0$ as one shows as in the proof for (1). Since $\hat{A} := |\hat{U}^T|\,\hat{D}\,|\hat{U}|$ is positive definite, then $\max\limits_{i,j} |\hat{a}_{ij}| = \max \hat{a}_{ii}$, and therefore

$$\|\hat{A}\|_\infty \leqq n \max_i \hat{a}_{ii}.$$

On the other hand, $\operatorname{diag} \hat{A} = \operatorname{diag}(\hat{U}^T \hat{D} \hat{U}) = \operatorname{diag}(A + E)$, hence

$$\hat{a}_{ii} = a_{ii} + e_{ii} \leqq a_{ii} + \gamma\,\hat{a}_{ii}$$

by (1′), and so $\max\limits_i \hat{a}_{ii} \leqq (1-\gamma)^{-1} \max\limits_i a_{ii}$. This shows that (with $\gamma \leqq 1/4$) the computed solution $\hat{x}$ to $Ax = b$ solves exactly

$$(A + \Delta A)\,\hat{x} = b \qquad \text{with } \|\Delta A\|_\infty \leqq 3\gamma n \max_i a_{ii} \leqq 3\gamma n \|A\|_\infty.$$

The customary analysis merely shows that the growth factor ϱ in (7) can be chosen to be 1 in this case.

The customary analysis and our analysis here are both based on the assumption that the rounding errors are small enough so that the pivotal elements (which would be positive in infinite precision arithmetic) remain positive. Wilkinson's much more detailed analysis [9] shows this to be the case provided $20nu\,n^{\frac{1}{2}}\|A\|_2\|A^{-1}\|_2 \leqq 1$. Such a result is possible since the Cholesky algorithm keeps the perturbation matrix E in (1′) symmetric and since all symmetric matrices in a small enough ball around a positive definite matrix are themselves positive definite.

Consider now the special case

$$|\hat{L}|\,|\hat{U}| = |\hat{L}\hat{U}|$$

which arises, e.g., when both $\hat{L}$ and $\hat{U}$ are nonnegative. We then have $|\hat{L}|\,|\hat{U}| = |\hat{L}\hat{U}| = |A + E| \leqq |A| + \gamma |\hat{L}|\,|\hat{U}|$, which implies

$$|\hat{L}|\,|\hat{U}| \leqq |A|/(1-\gamma).$$

On combining this inequality with (3), we obtain the following facts.

Proposition 1. *Assume that the $n \times n$ matrix A is invertible and has a factorization $A = LU$ with L unit lower triangular and U upper triangular, so that we may solve $Ax = b$ for x by Gauss elimination without pivoting. Assume now that the unit roundoff u is so small that $\gamma := n\,u/(1 - n\,u) \leqq 1/4$ and that we can compute the corresponding triangular factors $\hat{L}$ and $\hat{U}$ for A and then backsolve for $\hat{x}$. If $\hat{L}$ and $\hat{U}$ are nonnegative, then $\hat{x}$ solves exactly*

$$(A + \Delta A)\,\hat{x} = b \quad \textit{with } |\Delta A| \leqq 3\gamma |A|.$$

This is to be compared with (7) which gives the general estimate under the additional assumption that pivoting for size is used.

While the assumption $|\hat{L}|\,|\hat{U}| = |\hat{L}\hat{U}|$ is very special, it is usually satisfied when A is *totally positive*, i.e.,

$$\text{for all } r,\ i_1 < \dots < i_r,\ j_1 < \dots < j_r,\ \det A\begin{pmatrix} i_1, \dots, i_r \\ j_1, \dots, j_r \end{pmatrix} \geqq 0.$$

Proposition 2. *If A is totally positive and invertible, then A has a factorization $A = LU$ into a unit lower triangular matrix L and an upper triangular matrix U. Further, for $i > j$, $l_{ij} \geqq 0$ with equality if and only if $a_{pq} = 0$ for $p \geqq i$, $q \leqq j$. Also, for $i < j$, $u_{ij} \geqq 0$ with equality if and only if $a_{pq} = 0$ for $p \leqq i$, $q \geqq j$.*

Proof. Since A is totally positive, we have

$$0 \leqq \det A \leqq \det A\begin{pmatrix} 1, \dots, k \\ 1, \dots, k \end{pmatrix} \det A\begin{pmatrix} k+1, \dots, n \\ k+1, \dots, n \end{pmatrix}, \qquad k = 1, \dots, n$$

by Satz II.8 of [6] (or see Lemma 9.2 on p. 88 of [7]). The invertibility of A then implies that $\det A\begin{pmatrix} 1, \dots, k \\ 1, \dots, k \end{pmatrix} > 0$ for $k = 1, \dots, n-1$, which is a well known necessary and sufficient condition for the existence of the triangular factorization LU for A with L unit triangular (see, e.g., Theorem 9.2 of [5]). One finds easily that

$$\text{for } i > j,\ l_{ij} = \det A\begin{pmatrix} 1, \dots, j-1, i \\ 1, \dots, j-1, j \end{pmatrix} \Big/ \det A\begin{pmatrix} 1, \dots, j \\ 1, \dots, j \end{pmatrix}$$

hence $l_{ij} \geqq 0$ with equality iff $\det A\begin{pmatrix} 1, \dots, j-1, i \\ 1, \dots, j-1, j \end{pmatrix} = 0$. But, by the Hilfssatz 1 on p. 108 of [6] (used in the proof of Satz II.8) (or see 10.G on p. 96 of [7]), this will happen iff $a_{i1} = \dots = a_{ij} = 0$. Finally, if $a_{i1} = \dots = a_{ij} = 0$, then the invertibility of A implies that $a_{ir} \neq 0$ for some $r > j$, and then, by the total positivity of A, $a_{ir} > 0$ while

$$0 \leqq \det A\begin{pmatrix} i, p \\ q, r \end{pmatrix} = a_{iq}a_{pr} - a_{ir}a_{pq} = -a_{ir}a_{pq} \qquad \text{for } p > i \quad \text{and} \quad q \leqq j.$$

Thus also $a_{pq} = 0$ for $p > i$ and $q \leqq j$. The statement about the nonnegativity of U follows similarly. |||

This proposition allows the conclusion that, *in sufficiently high finite precision arithmetic, the computed triangular factors $\hat{L}$ and $\hat{U}$ for a totally positive invertible* *matrix A are nonnegative*. For, if $l_{ij}=0$ for some $i>j$, then $a_{pq}=0$, $p\geqq i$, $q\leqq j$ by the proposition and we must have $\hat{l}_{pq}=0$, for $p\geqq i$, $q\leqq j$ regardless of the size of the unit roundoff. Hence, L and $\hat{L}$ can only differ in the *nonzero* entries of L. Similarly, U and $\hat{U}$ only differ in the nonzero entries of U. But since $\hat{L}\to L$, $\hat{U}\to U$ as $u\to 0$ by (4), this implies that, for all sufficiently small unit roundoff u, all nonzero entries of $\hat{L}$ and $\hat{U}$ are actually positive.

We pointed out earlier Wilkinson's result which gives quantitative information in the positive definite case as to when the unit roundoff is small enough to make the error analysis applicable. A correspondingly informative statement in the totally positive case seems impossible since the property of being totally positive is much more delicate than that of being positive definite. The limit relation (4) can certainly be quantified in terms of the nonzero minors of A. But the result is neither pretty nor, we think, very useful since one is not likely to know, in practice, (lower bounds for) these minors. At the same time, the property on which our error analysis is based, i.e., the nonnegativity of the computed triangular factors, is easily monitored during the calculation.

Cryer [2, 3] has investigated the triangular factorization of totally positive matrices. Among his results is the remarkable fact that a matrix A is totally positive iff $A=LU$ for a totally positive lower triangular L and a totally positive upper triangular U, regardless of whether or not A is invertible.

3. Application to Spline Interpolation

Let $\boldsymbol{t}=(t_i)_1^{n+k}$ be nondecreasing with $t_i<t_{i+k-1}$, all i, and let $(N_i)_1^n$ be the corresponding sequence of B-splines of order k with knot sequence $\boldsymbol{t}$. Explicitly,

$$N_i(x)=(t_{i+k}-t_i)\,[t_i,\ldots,t_{i+k}]\,(\cdot-x)_+^{k-1},$$

i.e., $N_i(x)$ is $(t_{i+k}-t_i)$ times the k-th divided difference at $t_i,\ldots,t_{i+k}$ of the function $(t-x)_+^{k-1}:=(\max\{t-x,0\})^{k-1}$ of t. We denote by

$$\$_{k,\boldsymbol{t}}:=\left\{\sum_{i=1}^{n} a_i N_i \,\middle|\, \boldsymbol{a}\in\mathbb{R}^n\right\}$$

the collection of all *splines of order k with knot sequence $\boldsymbol{t}$*.

Consider the problem of interpolating to a given function g at the points $\tau_1<\ldots<\tau_n$ by elements of $\$_{k,\boldsymbol{t}}$. This leads to the linear system

$$\sum_{j=1}^{n} N_j(\tau_i)\,a_j=g(\tau_i),\qquad i=1,\ldots,n, \tag{8}$$

for the B-spline coefficients of the interpolating spline. Its coefficient matrix $(N_j(\tau_i))$ is invertible iff

$$N_i(\tau_i)\neq 0,\qquad i=1,\ldots,n, \tag{9}$$

according to the Schoenberg-Whitney theorem (see [1] for a simple proof).

Assume (9) to hold. Then, since

$$N_i=0\quad\text{off } (t_i,t_{i+k}),$$

we have

$$N_j(\tau_i)=0\quad\text{for } |i-j|\geqq k.$$

Hence $(N_j(\tau_i))$ is $(2k-1)$-banded. At the same time, $(N_j(\tau_i))$ is totally positive, by a theorem due to Karlin (e.g., Lemma 4.2 on p. 524 of [7]). We conclude that we can solve (8) by Gauss elimination *without pivoting*, hence without having to enlarge the bandwidth, and that the resulting computed solution $\hat{a}$ is the exact solution of a linear system

$$\sum_{j=1}^{n} \hat{a}_j (N_j(\tau_i) + e_{ij}) = g(\tau_i), \qquad i=1,\ldots,n \tag{10}$$

with

$$|e_{ij}| \leqq 3\,k\,u\,|N_j(\tau_i)|, \qquad \text{all } i,j, \tag{11}$$

provided $k\,u \leqq 0.01$. We stress the fact that *this bound does not depend on n.*

Similar results hold for the more general spline interpolation problem when $\boldsymbol{a}$ is to be determined so that

$$\int \varphi_i \sum_{j=1}^{n} a_j N_j = \int \varphi_i g, \qquad i=1,\ldots,n, \tag{12}$$

with $(\varphi_i)_1^n$ the B-splines of some order r and for some knot sequence $(\tau_i)_1^{n+r}$. For, the coefficient matrix in (12) is again banded, and is totally positive by Karlin's result and the Cauchy-Binet formula for the minors of a product. In particular, such results hold for least-squares approximation by splines.

Finally, we note that totally positive matrices arise in other interpolation processes, such as interpolation or least squares approximation by polynomials on $[0, a]$, by exponential sums and by certain rational functions. In particular, the Hilbert matrix is totally positive.

References

1. de Boor, C.: Total positivity of the spline collocation matrix. MRC TSR #1396, 1973. Ind. U. Math. J. **25**, 541–551 (1976)
2. Cryer, C.: The *LU*-factorization of totally positive matrices. Lin. Alg. Appl. **7**, 83–92 (1973)
3. Cryer, C.: Some properties of totally positive matrices. Comp. Sci. Tech. Rep. #207, U. Wisconsin, Madison, WI., Jan. 1974. Lin. Alg. Appl. (to appear)
4. Erisman, A. M., Reid, J. K.: Monitoring the stability of the triangular factorization of a sparse matrix, Numer. Math. **22**, 183–186 (1974)
5. Forsythe, G., Moler, C.: Computer solution of linear algebraic systems. Englewood Cliffs, N.Y.: Prentice-Hall 1967
6. Gantmacher, F. R., Krein, M. G.: Oszillationsmatrizen, Oszillationskerne und kleine Schwingungen mechanischer Systeme. Berlin: Akademie-Verlag 1960
7. Karlin, S.: Total positivity, Vol. I. Stanford, CA.: Stanford University Press 1968
8. Stoer, J.: Einführung in die Numerische Mathematik, Bd. I. Berlin-Heidelberg-New York: Springer 1972
9. Wilkinson, J. H.: A priori error analysis of algebraic processes. Proc. Internat. Congr. Math. (Moscow 1966), pp. 629–639. Moscow: Izdat. Mir 1968

Carl de Boor
Allan Pinkus
Mathematics Research Center
University of Wisconsin
610 Walnut Street
Madison, WI 53706, USA