SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 2, pp. 155–164, April 1989


003

# LARGE GROWTH FACTORS IN GAUSSIAN ELIMINATION WITH PIVOTING*

NICHOLAS J. HIGHAM† AND DESMOND J. HIGHAM‡

**Abstract.** The growth factor plays an important role in the error analysis of Gaussian elimination. It is well known that when partial pivoting or complete pivoting is used the growth factor is usually small, but it can be large. The examples of large growth usually quoted involve contrived matrices that are unlikely to occur in practice. We present real and complex $n \times n$ matrices arising from practical applications that, for any pivoting strategy, yield growth factors bounded below by $n/2$ and $n$, respectively. These matrices enable us to improve the known lower bounds on the largest possible growth factor in the case of complete pivoting. For partial pivoting, we classify the set of real matrices for which the growth factor is $2^{n-1}$. Finally, we show that large element growth does not necessarily lead to a large backward error in the solution of a particular linear system, and we comment on the practical implications of this result.

**Key words.** Gaussian elimination, growth factor, partial pivoting, complete pivoting, backward error analysis, stability

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

## 1. Introduction.

In his famous backward error analysis, Wilkinson proved that if the linear system $Ax = b$, where $A$ is $n \times n$, is solved in floating point arithmetic by Gaussian elimination with partial pivoting or complete pivoting, then the computed solution $\hat{x}$ satisfies (see, for example, [27, p. 108])

$$(A+E)\hat{x} = b, \tag{1.1a}$$

where

$$\|E\|_\infty \leq \rho_n p(n) u \|A\|_\infty. \tag{1.1b}$$

Here, $p(n)$ is a cubic polynomial in $n$, $u$ is the unit roundoff, and $\rho_n$ is the *growth factor*, defined in terms of the quantities $a_{ij}^{(k)}$ occurring during the elimination by

$$\rho_n = \rho_n(A) = \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|}.$$

As Wilkinson notes, the term $p(n)$ arises from bounds in the analysis that are rarely attained, and for practical purposes we can replace $p(n)$ by $n$ in (1.1b). Hence whether or not the bound in (1.1b) compares favourably with the "ideal" bound $\|E\|_\infty \leq u\|A\|_\infty$ depends on the size of the growth factor.

Although the growth factor is one of the most well-known quantities in numerical analysis, its behaviour when pivoting is used is not completely understood. Current

* Received by the editors May 16, 1988; accepted for publication (in revised form) September 9, 1988. The work of the second author was supported by a Science and Engineering Research Council Research Studentship.

† Department of Mathematics, University of Manchester, Manchester M13 9PL, United Kingdom. Present address, Department of Computer Science, Cornell University, Ithaca, New York 14853 (na.nhigham@na-net.stanford.edu).

‡ Department of Mathematics, University of Manchester, Manchester M13 9PL, United Kingdom. Present address, Department of Computer Science, University of Toronto, Toronto, Canada M5S 1A4 (na.dhigham@na-net.stanford.edu).

knowledge, in the context of general, dense matrices, can be summarised as follows. For clarity we will denote the growth factors for partial and complete pivoting by $\rho_n^p$ and $\rho_n^c$, respectively.

**Partial pivoting.** (The pivot element is selected as the element of largest absolute value in the active part of the pivot column.) The bound $\rho_n^p \leqq 2^{n-1}$ holds and is attained for matrices $A_n \in \mathbf{R}^{n\times n}$ of the following form [28, p. 212]:

$$A_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix}, \tag{1.2}$$

and also for $\tilde{A}_n = DA_nD$, $D = \operatorname{diag}(1, -1, 1, -1, \cdots, (-1)^{n+1})$ [26, p. 289]. Concerning the size of $\rho_n^p$ in practice, Wilkinson [28, pp. 213–214] says: "It is our experience that any substantial increase in size of elements of successive $A_r$ is extremely uncommon even with partial pivoting . . . No example which has arisen naturally has in my experience given an increase by a factor as large as 16." We are aware of no reports in the literature of experiences contrary to these related by Wilkinson over two decades ago. The largest growth factor that we have seen reported for a matrix not of the type (1.2) is $\rho_{100}^p = 35.1$, occurring for a symmetric matrix with elements from the uniform distribution on $[-1, 1]$ [18]; an earlier "record" value is $\rho_{40}^p = 23$, occurring for a random matrix of $1$s, $0$s and $-1$s [10, p. 1.21].

**Complete pivoting.** (The pivot element is selected as the element of largest absolute value in the whole of the remaining square submatrix.) Wilkinson [26, pp. 282–285] has shown that with complete pivoting

$$\rho_n^c \leqq n^{1/2}(2^{1}3^{1/2}4^{1/3}\cdots n^{1/(n-1)})^{1/2} \sim Cn^{1/2}n^{1/4\log n},$$

and that this bound is not attainable. He states in [26, p. 285] that "no matrix has been encountered in practice for which $p_1/p_n$ was as large as 8," and in [28, p. 213] that "no matrix has yet been discovered for which $f(r) > r$." ($p_i = (n - i + 1)$st pivot, $f(r) \equiv \rho_r^c$.)

Cryer [7] defines

$$g(n) = \sup_{A\in\mathbf{R}^{n\times n}} \rho_n^c(A). \tag{1.3}$$

The following results are known:

- $g(2) = 2$ (trivial).
- $g(3) = 2\frac{1}{4}$; Tornheim (see [7]) and Cohen [6].
- $g(4) = 4$; Cryer [7].
- $g(5) < 5.005$; Cohen [6].

Tornheim (see [7]) has shown that $\rho_n^c(H_n) \geqq n$ for any $n \times n$ Hadamard matrix $H_n$. $H_n$ is a Hadamard matrix if each $h_{ij} \in \{-1, 1\}$ and the rows of $H_n$ are mutually orthogonal. Hadamard matrices exist only for certain $n$; a necessary condition for their existence if $n > 2$ is that $n$ is a multiple of four. For more about Hadamard matrices see [14, Chap. 14] and [25].

Cryer [7] conjectured that for real matrices $\rho_n^c(A) \leqq n$, with equality if and only if $A$ is a Hadamard matrix. This conjecture is known to be false for complex matrices because Tornheim has constructed a $3 \times 3$ complex matrix $A$ for which $\rho_3^c(A) > 3$ (see [7]).

As the summary above indicates, most of what is known about the growth factor had been discovered by the early 1970s. Recently, Trefethen [23] has drawn attention to the shortcomings of our knowledge about the growth factor and asked, as one of his three mysteries, "Why is the growth of elements during elimination [with partial pivoting] negligible in practice?" Trefethen and Schreiber [24] have proposed a statistical analysis to explain why the growth factor is usually small for partial pivoting.

In this work we take a different approach from that of Trefethen and Schreiber. Instead of trying to explain small growth we pursue examples of large growth, and we investigate the implications of a large growth factor for numerical stability.

In § 2 we present several families of real matrices for which $\rho_n^c$ is bounded below by approximately $n/2$, and one family of complex matrices for which $\rho_n^c \geqq n$. Thus we obtain new lower bounds for $g(n)$ valid for all $n$. We also classify the real matrices for which $\rho_n^p = 2^{n-1}$, finding this to be a much richer class than might at first be thought.

In § 3 we reappraise the role of the growth factor in the backward error analysis of Gaussian elimination. We demonstrate that when solving linear systems by Gaussian elimination with partial pivoting large growth does not always induce a large backward error—there are certain, special right-hand sides for which the growth has no detrimental effect on the solution. We discuss the practical implications of this property for linear equation solvers.

**2. Matrices with a large growth factor.** We begin with a result that shows how to obtain a lower bound for the growth factor in Gaussian elimination. The bound applies whatever strategy is used for interchanging rows and columns, but we will be concerned only with partial and complete pivoting.

THEOREM 2.1. *Let $A \in \mathbf{C}^{n\times n}$ be nonsingular, and set $\alpha = \max_{i,j} |a_{ij}|$, $\beta = \max_{i,j} |(A^{-1})_{ij}|$, and $\theta = (\alpha\beta)^{-1}$. Then $\theta \leqq n$, and for any permutation matrices $P$ and $Q$ such that $PAQ$ has an LU factorisation, the growth factor $\rho_n$ for Gaussian elimination without pivoting on $PAQ$ satisfies $\rho_n \geqq \theta$.*

*Proof.* The inequality $\theta \leqq n$ follows from $\sum_{j=1}^n a_{ij}(A^{-1})_{ji} = 1$. Consider an LU factorisation $PAQ = LU$ computed by Gaussian elimination. We have

$$\begin{aligned} |u_{nn}^{-1}| &= |e_n^T U^{-1} e_n| = |e_n^T U^{-1} L^{-1} e_n| = |e_n^T Q^T A^{-1} P^T e_n| \\ &= |(A^{-1})_{ij}| \quad \text{for some } i,j \\ &\leqq \beta. \end{aligned}$$

Hence $\max_{i,j,k} |a_{ij}^{(k)}| \geqq |u_{nn}| \geqq \beta^{-1}$, and the result follows. □

*Remarks.* (1) $\theta^{-1} = \alpha\beta$ satisfies $\kappa_\infty(A)/n^2 \leqq \theta^{-1} \leqq \kappa_\infty(A)$, where the condition number $\kappa_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty$. Clearly, $A$ has to be very well-conditioned for the theorem to provide a lower bound $\theta$ near the maximum of $n$.

(2) In the case of partial pivoting $Q = I$, and the proof of Theorem 2.1 shows that we can take $\beta = \max_j |(A^{-1})_{nj}|$, which leads to a lower bound $\theta$ potentially larger than the one in the theorem.

(3) The relation $u_{nn}^{-1} = (A^{-1})_{ij}$ is used also in [4], with the aim of investigating cases where $u_{nn}$ is *small*.

To illustrate the theorem, consider a Hadamard matrix $H_n$. We have $H_n H_n^T = nI$, and so $H_n^{-1} = n^{-1}H_n^T$. Since $|h_{ij}| \equiv 1$, the theorem gives $\rho_n \geqq n$. As a special case we obtain $\rho_n^c(H_n) \geqq n$, as in [7] (this derivation is essentially the same as the one in [7]).

We present six further matrices to which the theorem can profitably be applied:

$$C_1 = \left(\cos\left(\frac{(i-1)(j-1)\pi}{n-1}\right)\right)_{i,j=1}^{n}, \qquad \rho_n \geqq \frac{n-1}{2}, \tag{2.1}$$

$$C_2 = \left(\cos\left(\frac{(i-1)(j-\frac{1}{2})\pi}{n}\right)\right)_{i,j=1}^{n}, \qquad \rho_n \geqq \frac{n}{2}, \tag{2.2}$$

$$S = \sqrt{\frac{2}{n+1}}\left(\sin\left(\frac{ij\pi}{n+1}\right)\right)_{i,j=1}^{n}, \qquad \rho_n \geqq \frac{n+1}{2}, \tag{2.3}$$

$$Q = \frac{2}{\sqrt{2n+1}}\left(\sin\left(\frac{2ij\pi}{2n+1}\right)\right)_{i,j=1}^{n}, \qquad \rho_n \geqq \frac{2n+1}{4}, \tag{2.4}$$

$$F = (f_{ij})_{i,j=1}^{n=2m}, \tag{2.5}$$

$$f_{ij} = \begin{cases} \cos\left(\dfrac{(i-1)(j-1)\pi}{m}\right), & 1 \leqq i \leqq m+1 \\ \sin\left(\dfrac{(i-m-1)(j-1)\pi}{m}\right), & m+2 \leqq i \leqq n \end{cases}, \qquad \rho_n \geqq \frac{n}{2},$$

$$V = \left(\exp\left(2\pi i(r-1)\frac{(s-1)}{n}\right)\right)_{r,s=1}^{n}, \qquad \rho_n \geqq n. \tag{2.6}$$

$C_1$ and $C_2$ are examples of Vandermonde-like matrices $C(\alpha_1, \alpha_2, \cdots, \alpha_n) = (T_{i-1}(\alpha_j))$ based on the Chebyshev polynomials $T_k$. (For further details of Vandermonde-like matrices and their applications see [16].) For $C_1$ the points $\alpha_j = \cos((j-1)\pi/(n-1))$ are the extrema of $T_{n-1}$, and for $C_2$ the points $\alpha_j = \cos((j-\frac{1}{2})\pi/n)$ are the zeros of $T_n$. The Chebyshev polynomials satisfy orthogonality conditions over both these sets of points [15, pp. 472–473]. Using these orthogonality properties, we can show that

$$C_1 D C_1 = \frac{(n-1)}{2} D^{-1}, \qquad D = \operatorname{diag}(\tfrac{1}{2}, 1, 1, \cdots, 1, \tfrac{1}{2}),$$

and

$$C_2 C_2^T = n \operatorname{diag}(1, \tfrac{1}{2}, \tfrac{1}{2}, \cdots, \tfrac{1}{2}).$$

Hence $C_1^{-1} = (2/(n-1))DC_1D$, and Theorem 2.1 yields $\rho_n(C_1) \geqq (n-1)/2$. It is not hard to show that for partial pivoting $u_{nn} = n-1$, and so $\rho_n^p(C_1) \geqq n-1$. Similarly, $C_2^{-1} = n^{-1}C_2^T \operatorname{diag}(1, 2, 2, \cdots, 2)$, and Theorem 2.1 gives $\rho_n(C_2) \geqq n/2$.

$S$ is the symmetric, orthogonal eigenvector matrix for the second difference matrix (the tridiagonal matrix with typical row $(-1, 2, -1)$) [22, p. 457]. Theorem 2.1 gives $\rho_n(S) \geqq (n+1)/2$. Another application in which $S$ and $C_2$ appear is the analysis of time series [1, § 6.5].

$Q$ is symmetric and orthogonal [19] and Theorem 2.1 yields $\rho_n(Q) \geqq (2n+1)/4$.

The matrix $F$, of even order $n = 2m$, arises in the derivation of approximations to linear operators for periodic functions. Hamming [15, pp. 522–524] shows that

$$F^{-1} = \frac{2}{n} F^T \operatorname{diag}(d_i), \qquad d_i = \begin{cases} \frac{1}{2}, & i = 1, m, \\ 1 & \text{otherwise.} \end{cases}$$

Hence Theorem 2.1 yields $\rho_n(F) \geqq n/2$.

Finally, $V$ is a complex Vandermonde matrix based on the roots of unity. It occurs in Fast Fourier Transform theory [22, pp. 292, 448]. $V^H V = nI$, so $V^{-1} = n^{-1}V^H$ and Theorem 2.1 gives $\rho_n(V) \geqq n$.

These matrices are not isolated examples: for each, the lower bound $\theta$ for $\rho_n$ is insensitive to small perturbations of the matrix. To see this, note that for $\|A^{-1}E\|_\infty < 1$ (say), in the notation of Theorem 2.1,

$$\theta(A+E)^{-1} = \alpha(A+E)\beta(A+E) \leqq (\alpha(A)+\alpha(E))(\beta(A)+O(\|E\|_\infty))$$
$$= \theta(A)^{-1}(1+O(\|E\|_\infty)).$$

Regarding the perturbation $E$ as the backward error in a computed LU factorisation, it follows also that, as long as $E$ is not too large, the computed growth factors will satisfy the theoretical lower bounds to within roundoff.

It is natural to ask what are the actual growth factors $\rho_n^p$ and $\rho_n^c$ for the matrices above. In numerical tests we found $\rho_n^p$ and $\rho_n^c$ generally to be bigger than the lower bounds, but appreciably less than $n$, except in the case of $V$ in (2.6) for which numerical evidence suggests that $\rho_n^p(V) = \rho_n^c(V) = n$.

All the above matrices are natural, noncontrived ones that arise in practical applications. For $n = 50$ (say), for both partial and complete pivoting, each of the matrices produces growth factors which exceed the generally accepted "maximum values in practice", such as the value 16 mentioned by Wilkinson in [28]. It is rather surprising that the growth factor properties of these examples have not previously been recognised. One possible explanation is that since each of the matrices is either an orthogonal or a diagonal scaling of an orthogonal matrix, Gaussian elimination may rarely have been applied to these matrices. (The growth factor properties of $C_1$ and $C_2$ were discovered incidentally when making a numerical comparison between Gaussian elimination with partial pivoting and a fast $O(n^2)$ algorithm [16].)

These examples provide new lower bounds for the maximum growth factor with complete pivoting. Specifically, we have, for $g(n)$ in (1.3),

$$g(n) \geqq \rho_n^c(S) \geqq \frac{n+1}{2} \quad \text{for all } n.$$

N. I. M. Gould (private communication) has suggested a way to obtain slightly sharper bounds: it is easy to show that

$$\theta(B) = \theta\left(\begin{bmatrix} A & A \\ A & -A \end{bmatrix}\right) = 2\theta(A),$$

and so, taking $A = S$, $g(2n) \geqq \rho_{2n}^c(B) \geqq \theta(B) = 2\theta(S) = n+1$, which improves on the lower bound $(2n+1)/2$. (Of course, for $n$ such that a Hadamard matrix $H_n$ exists, $g(n) \geqq n$ is a better bound; and for $n \leqq 5$ see the results quoted in § 1.) Furthermore, defining

$$\bar{g}(n) = \sup_{A \in \mathbf{C}^{n\times n}} \rho_n^c(A),$$

we have

$$\bar{g}(n) \geqq \rho_n^c(V) \geqq n.$$

The growth factors discussed above are relatively mild in the context of partial pivoting, since $O(n)$ growth falls significantly short of the potential $O(2^n)$. To investigate larger growth factors we have to make specific use of the properties of partial pivoting.

The following result shows that Wilkinson's example in which $\rho_n^p = 2^{n-1}$ is attained is just one from a nontrivial class of matrices with this property.

THEOREM 2.2. *All real $n \times n$ matrices $A$ for which $\rho_n^p(A) = 2^{n-1}$ are of the form*

$$A = DM\begin{bmatrix} T & \vdots & \theta d \\ 0 & \vdots & \end{bmatrix},$$

*where $D = \operatorname{diag}(\pm 1)$, $M$ is unit lower triangular with $m_{ij} = -1$ for $i > j$, $T$ is a nonsingular upper triangular matrix of order $n-1$, $d = (1, 2, 4, \cdots, 2^{n-1})^T$, and $\theta$ is a scalar such that $\theta = |a_{1n}| = \max_{i,j} |a_{ij}|$.*

*Proof.* Gaussian elimination with partial pivoting applied to a matrix $A$ gives a factorisation $B := PA = LU$, where $P$ is a permutation matrix. It is easy to show that $|u_{ij}| \leq 2^{i-1} \max_{r \leq i} |b_{rj}|$, with equality for $i = s$ only if there is equality for $i = 1, 2, \cdots, s-1$. Thus $\rho_n = 2^{n-1}$ implies that the last column of $U$ has the form $\theta Dd$, and also that $|b_{1n}| = \max_{i,j} |b_{ij}|$. By considering the final column of $B$, and imposing the requirement that $|l_{ij}| \leq 1$, it is easy to show that the unit lower triangular matrix $L$ must have the form $L = DMD$. It follows that at each stage of the reduction every multiplier is $\pm 1$; hence no interchanges are performed, that is, $P = I$. The only requirement on $T$ is that it be nonsingular, for if $t_{ii} = 0$ then the $i$th elimination stage would be skipped because of a zero pivot column, and no growth would be produced on that stage. □

In the case $n = 5$, the general form of $A$ is

$$A = D\begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} & \theta \\ -t_{11} & -t_{12}+t_{22} & -t_{13}+t_{23} & -t_{14}+t_{24} & \theta \\ -t_{11} & -t_{12}-t_{22} & -t_{13}-t_{23}+t_{33} & -t_{14}-t_{24}+t_{34} & \theta \\ -t_{11} & -t_{12}-t_{22} & -t_{13}-t_{23}-t_{33} & -t_{14}-t_{24}-t_{34}+t_{44} & \theta \\ -t_{11} & -t_{12}-t_{22} & -t_{13}-t_{23}-t_{33} & -t_{14}-t_{24}-t_{34}-t_{44} & \theta \end{bmatrix}.$$

We mention that it is straightforward to extend Theorem 2.2 to complex matrices.

As well as being of theoretical interest, the matrices given in this section are useful test matrices for linear equation solvers. Note that $\kappa_\infty(A)$ can be bounded above and below by multiples of $\kappa_\infty(T)$, so $T$ can be used to vary the condition of $A$. By varying the elements $m_{ij}$ $(i > j)$ and the vector $d$ in Theorem 2.2 we can construct matrices for which $\rho_n^p$ achieves any desired value between 1 and $2^{n-1}$. Indeed in practice it is expedient to modify $M$ in Theorem 2.2 so that $|m_{ij}| < 1$ for $i > j$, to ensure that rounding errors do not affect the pivot sequence (and hence the computed growth factor).

**3. Implications of a large growth factor.** If the growth factor $\rho_n$ is large then in the backward error result (1.1) the bound for $\|E\|_\infty$ is large. Whether or not $\|E\|_\infty$ itself is large when $\rho_n$ is large depends on the sharpness of the bound. Since the bound is independent of $b$, and $E$ clearly is not, we might suspect that the bound can be weak; in this section we will show that this is indeed the case.

We need to make use of an elementwise form of backward error analysis. Let $A \in \mathbb{R}^{n \times n}$. From [8] the computed solution $\hat{x}$ from Gaussian elimination (assuming, without loss of generality, no interchanges) satisfies

$$(A+F)\hat{x} = b, \tag{3.1a}$$

where

$$|F| \leq \gamma(2+\gamma)|\hat{L}|\,|\hat{U}|, \qquad \gamma = nu/(1-nu), \tag{3.1b}$$

and where $A \approx \hat{L}\hat{U}$ is the computed LU factorisation and $|F| = (|f_{ij}|)$.

As our use of the notation $E$ in (1.1) and $F$ in (3.1) suggests, the backward error for solution of $Ax = b$ is not uniquely defined: $G$ satisfying $(A + G)\hat{x} = b$ can be replaced by $G + H$ for any $H$ whose rows are orthogonal to $\hat{x}$. However, it is well known that of the infinitely many backward error matrices there is a unique one of minimal Frobenius norm,

$$G = \frac{r\hat{x}^T}{\hat{x}^T\hat{x}}, \qquad \|G\|_F = \frac{\|r\|_2}{\|\hat{x}\|_2}, \tag{3.2}$$

where the residual $r = b - A\hat{x}$, and $\|G\|_F = (\sum_{i,j} g_{ij}^2)^{1/2}$ (see [9, p. 171] for a proof and discussion, albeit in a different context). Of course, for the minimal Frobenius norm backward error matrix $G$ to be an appropriate one to consider, $A$ should be reasonably well-scaled.

Our aim is to obtain an informative bound for the minimal backward error $|G|$. To do this we write $r = b - A\hat{x} = F\hat{x}$, from (3.1a), and invoke the bound (3.1b), obtaining

$$|r| \leqq |F|\,|\hat{x}| \leqq \gamma(2+\gamma)|\hat{L}|\,|\hat{U}|\,|\hat{x}|.$$

Hence

$$|G| = \frac{|r|\,|\hat{x}|^T}{\|\hat{x}\|_2^2} \leqq \frac{\gamma(2+\gamma)}{\|\hat{x}\|_2^2}|\hat{L}|\,|\hat{U}|\,|\hat{x}|\,|\hat{x}|^T. \tag{3.3}$$

Our observation is that any large growth, which necessarily takes the form of large elements of $\hat{U}$ when partial pivoting is used, will not fully affect the backward error for a particular $\hat{x}$ if

$$\| |\hat{U}|\,|\hat{x}| \|_\infty \ll \|\hat{U}\|_\infty \|\hat{x}\|_\infty.$$

Since $|\hat{u}_{ij}| \leqq 2^{i-1} \max_{r \leq i} |a_{rj}|$, large growth can occur only toward the $(n, n)$ position of $\hat{U}$; consequently any $\hat{x}$ bounded by (say)

$$|\hat{x}| \leqq \|\hat{x}\|_\infty (1, 2^{-1}, 2^{-2}, \cdots, 2^{1-n})^T$$

can be shown to satisfy $\| |\hat{U}|\,|\hat{x}| \|_\infty \leqq 2\|A\|_\infty \|\hat{x}\|_\infty$, no matter how large $\|\hat{U}\|_\infty$.

For example, for any $A$, consider the use of partial pivoting for the particular system $Ax = b$ with $x = e_1$. Assume $\delta x = x - \hat{x}$ satisfies $\|\delta x\|_\infty \leqq 2^{1-n}/n$; this will certainly be the case if, making use of (1.1), $\kappa_\infty(A)4^{n-1}p(n)nu < 1$. Then

$$|\hat{L}|\,|\hat{U}|\,|\hat{x}| = |\hat{L}|\,|\hat{U}|\,|e_1 - \delta x| \leqq \max_r |a_{r1}|e + |\hat{L}|\,|\hat{U}|\,|\delta x|,$$

where $e = (1, 1, \cdots, 1)^T$, and thus

$$\| |\hat{L}|\,|\hat{U}|\,|\hat{x}| \|_\infty \leqq \|A\|_\infty + n2^{n-1}\|A\|_\infty \|\delta x\|_\infty \leqq 2\|A\|_\infty.$$

Hence, using (3.3), we have

$$\|G\|_\infty \leqq 2\gamma(2+\gamma)\|A\|_\infty(1 + O(2^{-n})),$$

which is an ideal backward error result, containing no growth factor term.

To illustrate the analysis we describe some numerical experiments performed using Gaussian elimination with partial pivoting and the perturbation $B_n = A_n + 0.1\,e_n e_n^T$ of Wilkinson's extreme growth matrix $A_n$ in (1.2). This perturbation of the $(n, n)$ element has the effect of causing rounding errors to be committed in the computation of the LU factorisation. Note that element growth occurs only in the *last* column of $B_n$ during Gaussian elimination with partial pivoting. For several $n$ we solved five different

linear systems $B_n x = b$, and computed the backward error for the LU factorisation, $\|B_n - \hat{L}\hat{U}\|_F / \|B_n\|_F$ (note that this is unique for a given norm), and the minimal backward error in the Frobenius norm for each system solved, $\|r\|_2/(\|B_n\|_F\|\hat{x}\|_2)$. For four of the linear systems, we selected $x$ or $b$ as vectors suggested by the analysis; for the final system we used a random $b$ with elements from the uniform distribution on [0, 1].

The computations were performed using the WATFOR-77 Fortran 77 compiler on a PC-AT compatible machine. Solutions were computed in single precision (IEEE standard, $u \approx 1.19 \times 10^{-7}$), using LINPACK's SGEFA/SGESL. The residuals $r$ and $B_n - \hat{L}\hat{U}$ were computed in double precision. The results are displayed in Table 3.1.

The backward errors for the LU factorisation are seen to be somewhat smaller than the large growth factor might lead us to expect, though still "alarmingly" large, except for $n = 10$. For $x = e_1$ the backward errors are all identically zero and $\hat{x} = x$; in this example the errors in the LU factorisation are nullified in the substitutions. The backward errors are also perfectly acceptable for $b = e_n$. Here the explanation is that $x_n = (B_n^{-1})_{nn} = u_{nn}^{-1}$, so that $u_{nn}x_n = 1$; thus the large elements in the last column of $\hat{U}$ vanish in the product $|\hat{U}|\,|\hat{x}|$ in (3.3). The backward errors for $b = e_1$, $b = e$, and the random $b$, all reflect the large backward error in the LU factorisation, as we would expect: the nonnegligible $x_n$ components pick out the large last column of $\hat{U}$ in the product $|\hat{U}|\,|\hat{x}|$.

To summarise, we have shown the following: When a linear system $Ax = b$ is solved by Gaussian elimination with partial pivoting, the backward error for the computed solution $\hat{x}$, $\|b - A\hat{x}\|_2/(\|A\|_F\|\hat{x}\|_2)$, can, in certain special cases, be substantially smaller than the backward error for the LU factorisation, $\|A - \hat{L}\hat{U}\|_F/\|A\|_F$, if the latter is large. Thus, strictly, the growth factor, or any other quantity appearing in a measure or bound of $A - \hat{L}\hat{U}$, is an unreliable indicator of the stability of a particular solution $\hat{x}$. We do not claim that this result is new, nor do we think that it will surprise anyone who has worked in backward error analysis. Examples of references that allude to the result in some way are [11, p. 73] and [20]. However we are not aware of a published analysis like the one above, and we feel that the result deserves to be better known.

It is important to stress that large growth is indeed very uncommon with partial pivoting (see the quotation from [28] in § 1), and that when it does occur there is a high probability that it will adversely affect the stability of the computed solution $\hat{x}$. Nevertheless, the result above has implications for how one uses a linear equation solver.

For example, consider the use of threshold versions of partial pivoting (including no pivoting at all); here large growth factors are much more common, and it is standard practice to monitor stability by estimating the error in the factorisation, $A - \hat{L}\hat{U}$ [5], [11]–[13]. If the estimate is large then a popular course of action is to carry out a

TABLE 3.1
*Results.*
($u \approx 1.19 \times 10^{-7}$)

| $n$ | $\rho_n^p$ | $\dfrac{\|B_n - \hat{L}\hat{U}\|_F}{\|B_n\|_F}$ | $\|r\|_2/(\|B_n\|_F\|\hat{x}\|_2)$ | | | | |
|---|---|---|---|---|---|---|---|
| | | | $z = e_1$ | $b = e_n$ | $b = e_1$ | $b = e$ | $b$ random |
| 10 | 4.7E2 | 3.0E-6 | 0.0 | 1.5E-8 | 3.1E-6 | 4.4E-6 | 2.3E-6 |
| 20 | 4.8E5 | 1.7E-3 | 0.0 | 5.5E-9 | 1.2E-3 | 1.6E-3 | 2.4E-4 |
| 30 | 4.9E8 | 4.5E-3 | 0.0 | 1.5E-11 | 3.2E-3 | 4.5E-3 | 1.1E-1 |
| 40 | 5.0E11 | 3.4E-3 | 0.0 | 1.1E-14 | 2.4E-3 | 3.4E-3 | 3.7E-2 |
| 50 | 5.1E14 | 2.7E-3 | 0.0 | 1.1E-17 | 1.9E-3 | 2.7E-3 | 4.4E-2 |
| 60 | 5.2E17 | 5.7E-2 | 0.0 | 1.5E-19 | 1.6E-3 | 2.3E-3 | 8.9E-2 |

refactorisation with a different pivot sequence. Our view is that if just a *single* system involving $A$ must be solved, it is worthwhile to proceed with the substitutions and to base refactorisation decisions on the easily computed *actual* backward error (3.2) rather than on (estimates of) $A - \hat{L}\hat{U}$, which may be misleading, as we have shown. For example, having computed $\hat{x}$ we might form $r = b - A\hat{x}$ (in single precision), evaluate the backward error $\|G\|_F = \|r\|_2 / \|\hat{x}\|_2$, and test whether $\|G\|_F \leqq \delta \|A\|_F$, where $\delta$ is an appropriate tolerance (depending on the unit roundoff, at least). Even if $\hat{x}$ is unacceptable, the substitutions need not have been wasted, for we may be able to achieve stability through the use of a few steps of iterative refinement [2], [3], [17], [21].

A more general way to express these views is that it is better to use *a posteriori* estimates that reflect the actual rounding errors encountered, rather than error estimates based on *a priori* analysis, such as (1.1). For a discussion of this philosophy we can do no better than refer the reader to Wilkinson's eloquent exposition in [29].

## REFERENCES


[1] T. W. ANDERSON, *The Statistical Analysis of Time Series*, John Wiley, New York, 1971.

[2] M. ARIOLI, J. W. DEMMEL, AND I. S. DUFF, *Solving sparse linear systems with sparse backward error*, Report CSS 214, Computer Science and Systems Division, Harwell Laboratory, AERE Harwell, Didcot, UK, 1988.

[3] Å. BJÖRCK, *Iterative refinement and reliable computing*, in Reliable Numerical Computation, M. G. Cox and S. J. Hammarling, eds., Oxford University Press, London, 1989.

[4] T. F. CHAN, *On the existence and computation of* LU*-factorizations with small pivots*, Math. Comp., 42 (1985), pp. 535–547.

[5] E. CHU AND A. GEORGE, *A note on estimating the error in Gaussian elimination without pivoting*, ACM SIGNUM Newsletter, 20 (1985), pp. 2–7.

[6] A. M. COHEN, *A note on pivot size in Gaussian elimination*, Linear Algebra Appl., 8 (1974), pp. 361–368.

[7] C. W. CRYER, *Pivot size in Gaussian elimination*, Numer. Math., 12 (1968), pp. 335–345.

[8] C. DE BOOR AND A. PINKUS, *Backward error analysis for totally positive linear systems*, Numer. Math., 27 (1977), pp. 485–490.

[9] J. E. DENNIS, JR. AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[10] J. J. DONGARRA, J. R. BUNCH, C. B. MOLER, AND G. W. STEWART, LINPACK *Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

[11] I. S. DUFF, A. M. ERISMAN, AND J. K. REID, *Direct Methods for Sparse Matrices*, Oxford University Press, London, 1986.

[12] A. M. ERISMAN, R. G. GRIMES, J. G. LEWIS, W. G. POOLE, AND H. D. SIMON, *Evaluation of orderings for unsymmetric sparse matrices*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 600–624.

[13] A. M. ERISMAN AND J. K. REID, *Monitoring the stability of the triangular factorization of a sparse matrix*, Numer. Math., 22 (1974), pp. 183–186.

[14] M. HALL, JR., *Combinatorial Theory*, Blaisdell, Waltham, MA, 1967.

[15] R. W. HAMMING, *Numerical Methods for Scientists and Engineers*, 2nd ed., McGraw-Hill, New York, 1973.

[16] N. J. HIGHAM, *Stability analysis of algorithms for solving confluent Vandermonde-like systems*, Numerical Analysis Report 148, University of Manchester, Manchester, UK, 1987.

[17] M. JANKOWSKI AND H. WOŹNIAKOWSKI, *Iterative refinement implies numerical stability*, BIT, 17 (1977), pp. 303–311.

[18] A. J. MACLEOD, *The distribution of the growth factor in Gaussian elimination with partial pivoting*, Technical Report, Department of Mathematics and Statistics, Paisley College of Technology, Paisley, Scotland, 1988.

[19] R. B. POTTS, *Symmetric square roots of the finite identity matrix*, Utilitas Math., 9 (1976), pp. 73–86.

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

[21] ———, *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comp., 35 (1980), pp. 817–832.

[22] G. STRANG, *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, Wellesley, MA, 1986.
[23] L. N. TREFETHEN, *Three mysteries of Gaussian elimination*, ACM SIGNUM Newsletter, 20 (1985), pp. 2–5.
[24] L. N. TREFETHEN AND R. S. SCHREIBER, *Average-case stability of Gaussian elimination*, Numerical Analysis Report 88-3, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, 1987; SIAM J. Matrix Anal. Appl., to appear.
[25] W. D. WALLIS, A. P. STREET, AND J. S. WALLIS, *Combinatorics: Room Squares, Sum-Free Sets, Hadamard Matrices*, Lecture Notes in Mathematics 292, Springer-Verlag, Berlin, 1972.
[26] J. H. WILKINSON, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281–330.
[27] ———, *Rounding Errors in Algebraic Processes*, Notes on Applied Science No. 32, Her Majesty's Stationery Office, London, 1963.
[28] ———, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.
[29] ———, *Error analysis revisited*, IMA Bulletin, 22 (1987), pp. 192–200.