MATHEMATICS OF COMPUTATION, VOLUME 35, NUMBER 151
JULY 1980, PAGES 817–832

Iterative Refinement Implies Numerical Stability for Gaussian Elimination

By Robert D. Skeel*

Abstract. Because of scaling problems, Gaussian elimination with pivoting is not always as accurate as one might reasonably expect. It is shown that even a single iteration of iterative refinement in single precision is enough to make Gaussian elimination stable in a very strong sense. Also, it is shown that without iterative refinement row pivoting is inferior to column pivoting in situations where the norm of the residual is important.

1. Introduction. It is well known that Gaussian elimination with pivoting is a stable algorithm for solving linear systems of equations in the sense that the computed solution exactly satisfies a linear system whose coefficient matrix differs slightly in norm from the given matrix. For this reason it is often thought that iterative refinement is not worthwhile unless either the data are known with great accuracy or one wishes to detect ill-conditioning. However, it has been pointed out (Hamming (1971), Gear (1975)) that Gaussian elimination is not as accurate as one might reasonably expect in that the computed solution may not exactly satisfy a linear system with *each coefficient* slightly different from that given. It is shown in Skeel (1979) that stability in this strong sense is possible if an appropriate implicit scaling of the rows and/or columns is used with the pivoting. Unfortunately the proper scaling requires estimates of the solution components. It is the purpose of this paper to show that the effects of improper scaling can be eliminated by performing iterative refinement even if the residuals are not accumulated in double precision. Therefore, iterative refinement would be worthwhile for problems that may not be scaled properly for Gaussian elimination. The computational cost is often small, but this is not always true due to the necessity of storing the original matrix.

The title of this paper is adapted from a related paper of Jankowski and Woźniakowski (1977). The principal result of their paper is that almost any linear equation solver can be made stable in the usual sense by performing iterative refinement

Received July 24, 1978; revised June 12, 1979.
1980 *Mathematics Subject Classification.* Primary 65F05.
Key words and phrases. Iterative refinement, iterative improvement, numerical stability, Gaussian elimination, pivoting, backward error analysis, roundoff analysis.
*Research sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR-75-2854. The United States Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation hereon.

0025-5718/80/0000-0111/$05.00

even if single precision is used throughout. In contrast, our paper shows that a particular algorithm which is already stable in the usual sense becomes stable in the strong sense when iterative refinement is performed. These latter results are obtained primarily through the use of componentwise absolute values instead of norms for the roundoff error analysis.

In Section 2 three stability concepts are defined which are motivated by the desire to compare the solution computed by the algorithm with solutions obtained by introducing small errors in each of the entries of the coefficient matrix.

In Section 3 a good error bound is given and numerical stability is discussed for Gaussian elimination with column pivoting, which is the usual variant of partial pivoting in which the largest element of the next column is used for a pivot.

In Section 4 bounds on the residual and the error are obtained for column pivoting with iterative refinement. Both s.p.r.a. (single precision residual accumulation) and d.p.r.a. (double precision residual accumulation) are considered.

In Section 5 the numerical stability of iterative refinement is examined, and it is shown that a single iteration in single precision is enough for stability. Still better behavior is possible with further iterations or with double precision residuals.

In Section 6 an error bound is given for Gaussian elimination with row pivoting, which means that columns are interchanged so that each pivot is the largest in its row. The interesting result here is that without iterative refinement the norm of the residual can be much larger than for column pivoting; otherwise, little can be said about the relative merits of the two types of partial pivoting.

2. Numerical Stability. Three stability concepts are discussed for algorithms that solve a system $Ax = b$ of n equations in n unknowns.

A floating-point number system consists of a subset of the reals for which floating-point operations $\hat{+}$, $\hat{-}$, $\hat{\times}$, and $\hat{/}$ are defined. It is assumed that the relative roundoff error of floating-point arithmetic is bounded by a minuscule positive number u satisfying the restriction $nu \leqslant .01$ of Forsythe and Moler (1967). Every reference to a floating-point result $x\hat{o}y$ carries with it the assumption that x, $\hat{o}$, and y are such that the result is well defined.

For a roundoff error analysis it is helpful to have some reasonable standard for comparison. As our standard, we would like to consider the slightly perturbed solution of a slightly perturbed problem; more specifically, errors of relative size $\leqslant \epsilon$ are introduced into the problem data, and then errors of relative size $\leqslant \epsilon$ are introduced into the exact solution of the perturbed problem. For the particular problem of solving a linear system $Ax = b$ there are indications (see Skeel (1979)) that it makes little difference if only the matrix A is perturbed, and so we consider solutions $x + \delta x = (A + \delta A)^{-1}b$ for δA smaller than A by a factor of ϵ. It is usual to consider δA such that $\|\delta A\| \leqslant \epsilon\|A\|$, but it may be preferable (see Skeel (1979)) to be more restrictive by requiring that $|\delta A| \leqslant \epsilon|A|$, where the inequality and the absolute value are to be understood in a componentwise sense.

There are various ways of relating the computed solution $\hat{x}$ to the solutions $x + \delta x$ of perturbed problems. One way is the backward error η, which is defined to be the

least amount by which A must be perturbed to get a solution $x + \delta x$ which is exactly equal to $\hat{x}$. The requirement that $x + \delta x = \hat{x}$ can be quite demanding, and so we also wish to consider less stringent conditions. We can ask instead only that $x + \delta x$ be in some sense as bad a solution as $\hat{x}$. We consider two measures of the badness of the solution: the norm of the residual and the norm of the error.

Definition. Let $\Delta_\epsilon(A, b) = \{\delta x: (A + \delta A)(x + \delta x) = b \text{ where } |\delta A| \leqslant \epsilon |A|\}$. The backward errors are defined by

$$\eta = \inf\{\epsilon: \hat{x} = x + \delta x \text{ for some } \delta x \in \Delta_\epsilon(A, b)\},$$
$$\eta^R = \inf\{\epsilon: \|A\hat{x} - b\| = \|A\delta x\| \text{ for some } \delta x \in \Delta_\epsilon(A, b)\},$$
$$\eta^E = \inf\{\epsilon: \|\hat{x} - x\| = \|\delta x\| \text{ for some } \delta x \in \Delta_\epsilon(A, b)\}.$$

Only absolute norms will be considered, namely those for which

$$\| |v| \| = \|v\|$$

for any vector v. Bauer, Stoer, and Witzgall (1961) show that this property is equivalent to monotonicity, which means that

$$\|v\| \leqslant \|w\| \quad \text{whenever } |v| \leqslant |w|.$$

In addition, it is convenient to assume that the problem has been scaled so that the norm can be chosen so that $\|e_j\| = 1$, $1 \leqslant j \leqslant n$. (Extension of the results to weighted norms would be straightforward.) Under these assumptions it is easily shown that $\|v\|_\infty \leqslant \|v\| \leqslant \|v\|_1$ for any vector v. The norm is to be extended to row vectors and matrices in the usual way.

By stability of an algorithm it is meant that there exists a stability constant $k(n)$ and a stability threshold $\bar{u}(n)$ such that the backward error $\eta \leqslant k(n)u$ whenever $u \leqslant \bar{u}(n)$. However, this definition is difficult to apply, and so we relax it by allowing $\bar{u}(n)$ to depend on the data (A, b) (cf. Jankowski and Woźniakowski (1977) and the asymptotic backward stability of Miller (1972)). An algorithm will be called R-stable if there exists a constant $k(n)$ such that $\eta^R \leqslant k(n)u$ for sufficiently small u and E-stable if $\eta^E \leqslant k(n)u$. Since both η^R and η^E are $\leqslant \eta$, it is clear that stability implies both R- and E-stability.

The following theorem gives in terms of $\hat{x}$ an expression for η which is useful both theoretically and computationally.

THEOREM 2.1 (OETTLI AND PRAGER, (1964)). *The backward error η of the computed solution $\hat{x}$ satisfies*

$$\eta = \max \frac{|A\hat{x} - b|}{|A||\hat{x}|}$$

if $|A||\hat{x}| > 0$, where division of two vectors is defined componentwise.

Proof. This is similar to Eq. (4.2) of Oettli and Prager (1964). A slightly stronger result without the hypothesis $|A||\hat{x}| > 0$ is proved in Skeel (1979). □

It does not seem possible to obtain similar expressions for η^R and η^E; however,

the next two theorems give excellent upper and lower bounds which are useful theoretically. These bounds contain the quantities

$$\kappa(A^{-1}) = \| \, |A| \, |A^{-1}| \, \| \quad \text{and} \quad \kappa(A) = \| \, |A^{-1}| \, |A| \, \|.$$

It is important to note that κ is being used in a nonstandard way, for it does not denote a condition number defined in terms of norms.

THEOREM 2.2. *The backward error η^R of the computed solution $\hat{x}$ satisfies*

$$\frac{\|A\hat{x} - b\|}{\| \, |A| \, |x| \, \| + \kappa(A^{-1})\|A\hat{x} - b\|} \leqslant \eta^R \leqslant \frac{\|A\hat{x} - b\|}{\| \, |A| \, |x| \, \| - \kappa(A^{-1})\|A\hat{x} - b\|}$$

provided that the denominator is positive; alternatively,

$$\eta^R \leqslant \frac{\|A\hat{x} - b\|}{\| \, |A| \, |x| \, \|_\infty - \|A\hat{x} - b\|}$$

provided that the denominator is positive.

Proof of Lower Bound. Let ϵ be any real such that $\|A\hat{x} - b\| = \|A\delta x\|$ for some δx which satisfies $(A + \delta A)(x + \delta x) = b$ where $|\delta A| \leqslant \epsilon |A|$. Then

$$\|A\delta x\| = \| - \delta A x - \delta A A^{-1} A \delta x \| \leqslant \epsilon \{ \| \, |A| \, |x| \, \| + \kappa(A^{-1}) \|A\delta x\| \},$$

which provides a lower bound on ϵ and hence on η^R. □

Proof of First Upper Bound. For $\epsilon \geqslant 0$ define

$$\delta A = \epsilon |A| \operatorname{diag}(\operatorname{sgn}(x))$$

and define $\delta x = (A + \delta A)^{-1} b - x$ so that δx is a rational function of ϵ where removable singularities are assumed to be removed. Hence, $\|A\delta x\|$ is a continuous function of ϵ except at poles of δx where $\|A\delta x\| = +\infty$. For values of ϵ not equal to poles of δx it can be shown that $(A + \delta A)(x + \delta x) = b$ (although it may not be true that $A + \delta A$ is nonsingular). Equivalently

$$\delta A x = -(I + \delta A A^{-1}) A \delta x,$$

whence

$$\|\delta A x\| \leqslant \|I + \epsilon |A| \, |A^{-1}| \, \| \, \|A\delta x\|$$

and

$$\|A\delta x\| \geqslant \frac{\epsilon \| \, |A| \, |x| \, \|}{1 + \epsilon \kappa(A^{-1})}. \tag{2.1}$$

Hence for $\epsilon \geqslant 0$, $\|A\delta x\|$ assumes all nonnegative values less than $\| \, |A| \, |x| \, \| / \kappa(A^{-1})$. By assumption the norm of the residual is less than this value and, therefore, choose ϵ so that $\|A\delta x\| = \|A\hat{x} - b\|$. Since ϵ cannot be a pole of δx, we have $(A + \delta A)(x + \delta x) = b$; and since $|\delta A| \leqslant \epsilon |A|$, it follows that $\eta^R \leqslant \epsilon$. Solving (2.1) for ϵ and substituting $\|A\hat{x} - b\|$ for $\|A\delta x\|$ establishes the bound. □

Proof of the Second Upper Bound. For $\epsilon \geqslant 0$ define

$$\delta A = \epsilon(1 + 2\epsilon)^{-1}A(\epsilon I + (1 + \epsilon)E),$$

where

$$E = \operatorname{diag}(\operatorname{sgn}(e_l^T A))\ \operatorname{diag}(\operatorname{sgn}(x));$$

and l is such that $e_l^T |A|\,|x| = \| \,|A|\,|x|\,\|_\infty$. Hence,

$$(A + \delta A)^{-1} = (I - \epsilon(1 + \epsilon)^{-1}E)A^{-1} \quad \text{and} \quad \delta x = (A + \delta A)^{-1}b - x = -\epsilon(1 + \epsilon)^{-1}Ex,$$

whence the residual $A\delta x = -\epsilon(1 + \epsilon)^{-1}AEx$. Thus,

$$|e_i^T A\delta x| \leqslant \epsilon(1 + \epsilon)^{-1}\,\| \,|A|\,|x|\,\|_\infty$$

with equality for $i = l$, and so

$$\|A\delta x\| \geqslant \|A\delta x\|_\infty = \epsilon(1 + \epsilon)^{-1}\,\| \,|A|\,|x|\,\|_\infty. \tag{2.2}$$

Therefore, $\|A\delta x\|$ assumes all nonnegative values less than $\| \,|A|\,|x|\,\|_\infty$; and so we may choose ϵ so that $\|A\delta x\| = \|A\hat{x} - b\|$. Since $(A + \delta A)(x + \delta x) = b$ and $|\delta A| \leqslant \epsilon|A|$, it follows that $\eta^R \leqslant \epsilon$. Solving (2.2) for ϵ and substituting $\|A\hat{x} - b\|$ for $\|A\delta x\|$ establishes the bound. □

THEOREM 2.3. *The backward error η^E of the computed solution $\hat{x}$ satisfies*

$$\frac{\|\hat{x} - x\|}{\| \,|A^{-1}|\,|A|\,|x|\,\| + \kappa(A)\|\hat{x} - x\|} \leqslant \eta^E \leqslant \frac{\|\hat{x} - x\|}{\| \,|A^{-1}|\,|A|\,|x|\,\|_\infty - \kappa(A)\|\hat{x} - x\|}$$

provided that the denominator is positive.

Proof of Lower Bound. Let ϵ be any real such that $\|\hat{x} - x\| = \|\delta x\|$ for some δx which satisfies $(A + \delta A)(x + \delta x) = b$ where $|\delta A| \leqslant \epsilon|A|$. Then

$$\|\delta x\| = \| -A^{-1}\delta Ax - A^{-1}\delta A\delta x\| \leqslant \epsilon\{\,\| \,|A^{-1}|\,|A|\,|x|\,\| + \kappa(A)\|\delta x\|\,\},$$

which provides a lower bound on ϵ and hence on η^E. □

Proof of Upper Bound. For $\epsilon \geqslant 0$ define

$$\delta A = \epsilon\ \operatorname{diag}(\operatorname{sgn}(e_l^T A^{-1}))|A|\ \operatorname{diag}(\operatorname{sgn}(x)),$$

where l is such that

$$e_l^T |A^{-1}|\,|A|\,|x| = \| \,|A^{-1}|\,|A|\,|x|\,\|_\infty;$$

and define $\delta x = (A + \delta A)^{-1}b - x$ so that δx is a rational function of ϵ where removable singularities are assumed to be removed. If ϵ is not a pole of δx, it can be shown that $(A + \delta A)(x + \delta x) = b$. Equivalently

$$A^{-1}\delta Ax = -(I + A^{-1}\delta A)\delta x,$$

whence

$$|e_i^T A^{-1}\delta A x| \leqslant \|I + \epsilon|A^{-1}||A|\|\,\|\delta x\| \quad \text{and} \quad \|\delta x\| \geqslant \frac{\epsilon\||A^{-1}||A||x|\|_\infty}{1+\epsilon\kappa(A)}.$$

The remainder of the proof is similar to the proof of the first upper bound of Theorem 2.2. □

Remark. The norm $\|\circ\|_\infty$ in the upper bound cannot be replaced by $\|\circ\|$; however, the two norms differ by at most a factor of n, which is not too important for our purposes.

From these theorems it is clear that R-stability is equivalent to

$$\|A\hat{x} - b\| \leqslant k(n)u\||A||x|\| + O(u^2)$$

and E-stability is equivalent to

$$\|\hat{x} - x\| \leqslant k(n)u\||A^{-1}||A||x|\| + O(u^2).$$

If absolute values were replaced by norms, then R- and E-stability would be equivalent to the good behavior and the stability, respectively, of Jankowski and Woźniakowski (1977).

In Skeel (1979) it is shown that $\||A^{-1}||A||x|\|_\infty/\|x\|_\infty$ is the condition number for the maximum norm of the error with respect to small relative changes in the elements of A. Similarly, it can be shown that $\||A||x|\|/\|Ax\|$ is the condition number for the norm of the residual (which is just the error measured with the norm $\|A(\circ)\|$) if the relative residual is defined to be $\|A\hat{x} - b\|/\|b\|$ (cf. Bauer (1963)). Moreover, $\kappa(A)$ and $\kappa(A^{-1})$ are upper bounds on the condition numbers for the error and the residual, respectively.

3. Gaussian Elimination With Column Pivoting. Error bounds are given and numerical stability is discussed for Gaussian elimination with column pivoting.

The remainder of this paper is limited to the consideration of fully *a priori* error bounds in terms of the problem data A and b. Thus, for example, the factor for the growth of elements in the elimination is undesirable because it depends in a very complicated way on the data and inappropriate because it depends on the details of the floating-point arithmetic. The bounds we seek are generally not computationally useful because they are realistic only for the worst case errors which may be many orders of magnitude greater than the typical errors. Nevertheless, such bounds seem to provide useful theoretical information, particularly for the purpose of comparing the stability of different algorithms.

Expressions for roundoff error bounds tend to be quite complicated, and the amount of detail can be overwhelming. Moreover, such bounds are often extremely pessimistic, and so it would seem appropriate to suppress some of the less relevant detail. We choose to conceal somewhat the specific functional dependence of various quantities on n. To accomplish this, we introduce symbols C_1, c_1, C_2, c_3, c_4, c_5, C_6, . . . , to represent positive quantities which are bounded above by functions of n only. The lower and upper case symbols represent scalars and matrices, respectively. These sym-

bols are all defined in paragraphs with the heading "Note," which the reader is encouraged to skip over. The reasons for providing these definitions are to establish the correctness of the results and to enable the interested reader to construct more detailed error bounds. There are two justifications for suppressing the functional dependence on n. The practical reason is that the function-of-n part of the error bound is the factor which is most unrealistic for typical cases. For example, the error bound for Gaussian elimination with partial pivoting contains the factor 2^n, and this bound is attained by examples like that of Wilkinson (1963). Nevertheless, it is observed (Forsythe, Malcolm, and Moler (1977, p. 46)) that in practice the error is bounded independently of n with rare exceptions; and for this reason most authors avoid stating explicit a priori bounds for partial pivoting. The theoretical reason is that the stability concepts of roundoff analysis require the existence of bounds which do not depend on A and b but may depend on n. (Each value of n corresponds to a different function $f(A, b) = A^{-1}b$.) Thus, knowledge of the dependence on A and b is crucial to establishing stability results, but dependence on n is irrelevant.

The basic result, shown in Skeel (1979), upon which our error analysis depends, is that Gaussian elimination with column pivoting determines an approximation $\hat{x}$ to $x = A^{-1}b$ which satisfies

$$|A\hat{x} - b| \leqslant uC_1|A||\hat{x}|$$

for some nonnegative matrix C_1 depending only on n. In fact, C_1 is a lower triangular matrix with its rows permuted, and $\|C_1\|_\infty \leqslant \lceil 19 \cdot 2^{n-2} - n - 8 \rceil e^{2nu}$ under certain assumptions on the minor details of the computation. This bound also holds for complete pivoting with a much smaller value for C_1. The same would be true for a column pivoting algorithm which monitors the element growth and switches to complete pivoting if the growth factor exceeds some predetermined threshold value.

The error bound is not quite an a priori bound because it contains the computed quantity $\hat{x}$. This can be eliminated by writing

$$\begin{aligned} |A\hat{x} - b| &\leqslant uC_1|A||x + A^{-1}(A\hat{x} - b)| \\ &\leqslant uC_1|A||x| + uC_1|A||A^{-1}||A\hat{x} - b| \end{aligned}$$

and solving for $|A\hat{x} - b|$ to get

$$|A\hat{x} - b| \leqslant uC_2|A||x|, \tag{3.1}$$

assuming that $c_1 u\kappa(A^{-1}) \leqslant \tfrac{1}{2}$.

Note. The quantities

$$c_1 = \|C_1\|, \quad C_2 = (I - uC_1|A||A^{-1}|)^{-1}C_1$$

are bounded above by functions of n only.

Backward error bounds for Gaussian elimination follow from Theorems 2.1, 2.2, and 2.3. For example, the following stability bounds for Gaussian elimination can be obtained:

$$\eta \leqslant c_1 u \max \frac{e\||A||x|\|}{|A||x|} + O(u^2),$$

where $e = (1, 1, \ldots, 1)^T$,

$$\eta^R \leqslant c_1 u + O(u^2), \quad \eta^E \leqslant c_1 u \frac{\|A^{-1}\| \| |A| |x| \|}{\| |A^{-1}| |A| |x| \|} + O(u^2).$$

Hence partial pivoting is R-stable, which is consistent with the observation of Forsythe, Malcolm, and Moler (1977) that "It is probably the single most important fact which people concerned with matrix computations have learned in the past 15 or 20 years: *Gaussian elimination with partial pivoting is guaranteed to produce small residuals.*" However, examples exist (Hamming (1971), p. 120)) showing that partial pivoting is not E-stable (and, therefore, not stable), although Skeel (1979) shows that the quantities

$$\sigma(A, x) = \max \frac{e\| |A| |x| \|}{|A| |x|} \quad \text{and} \quad \tau(A, x) = \frac{\|A^{-1}\| \| |A| |x| \|}{\| |A^{-1}| |A| |x| \|}$$

can be reduced to unity by the use of an appropriate scaling of the rows in conjunction with partial pivoting. Thus, $\sigma(A, x)$ and $\tau(A, x)$ are measures of ill-scaling for the system of equations. The proper scaling, however, cannot be determined efficiently, and it is the purpose of this paper to show that iterative refinement eliminates the effects of poor scaling, thus making Gaussian elimination numerically stable.

4. Error Bounds. We obtain bounds on the residual and the error for each iteration of iterative refinement in terms of A, b, u and anonymous quantities bounded above by functions of n only. (Very detailed bounds are given in Skeel (1977).) Both single and double precision accumulation of the residuals are considered.

Iterative refinement is defined as follows, where subscripts denote iterates rather than components of vectors:

$$x_1 = \text{value of } A^{-1}b \text{ computed by column pivoting,}$$

for $m = 1, 2, \ldots,$

$$r_m = \text{computed value of } Ax_m - b,$$
$$d_m = \text{value of } A^{-1}r_m \text{ computed by column pivoting,}$$
$$x_{m+1} = x_m \hat{-} d_m.$$

It is also convenient to define

$$x_0 = 0, \quad r_0 = -b, \quad d_0 = -x_1.$$

The residual r_m is to be computed with the subtraction performed last. For s.p.r.a. the computation is done in single precision and for d.p.r.a. the computation is done in double precision followed by a conversion of the result to single precision. It is assumed that the relative roundoff error of double precision arithmetic is bounded by u^2 and that the relative error of conversion is bounded by u.

It is quite easy to show that Gaussian elimination followed by one refinement in single precision is a stable algorithm according to our definition of stability. First, the computed solution x_1 satisfies $Ax_1 = b + g_0$, where $g_0 = O(u)$. Second, the

computed residual

$$r_1 = Ax_1 - b + f_1 = g_0 + f_1,$$

where

$$|f_1| \leqslant nu|A||x_1| + u|Ax_1 - b| + O(u^2) = nu|A||x| + O(u^2).$$

Third, the computed error d_1 satisfies

$$|Ad_1 - r_1| \leqslant uC_1|A||d_1| = O(u^2),$$

and hence,

$$Ad_1 = g_0 + f_1 + O(u^2).$$

Fourth, the refined solution

$$x_2 = x_1 - d_1 + h_2,$$

where $|h_2| \leqslant u|x_1 - d_1| = u|x| + O(u^2)$. Thus,

$$Ax_2 - b = -f_1 + Ah_2 + O(u^2) \quad \text{and} \quad |Ax_2 - b| \leqslant (n+1)u|A||x| + O(u^2),$$

which together with $|A||x_2| = |A||x| + O(u)$ implies stability according to Theorem 2.1. Miller and Wrathall (1979) note that this result is true *even without pivoting*, but Miller (1977) suggests that it tells us much more about asymptotic notions of stability than about Gaussian elimination. Therefore, we perform a more detailed analysis that indicates the size of the stability threshold $\bar{u}(n)$.

The error analysis that follows is quite laborious, and there are three reasons for this:

(i) *generality.* Results are obtained for any number of iterations with either s.p.r.a. or d.p.r.a., and they are applicable to any of the three stability measures with any absolute norm.

(ii) *sharpness.* The lemmas and Theorem 4.4 seem to give the best possible bounds of the type we seek.

(iii) *adaptability.* The analysis is easily modified so that it applies to row pivoting instead of column pivoting.

The reader may wish to skip to the discussion following Theorem 4.4.

Error bounds for an iterative process are usually obtained by bounding the $(m+1)$th iterate of some quantity in terms of the mth iterate of this quantity. The quantity which is selected for this purpose affects the sharpness of the results; a good choice seems to be the exact residual of $x_m - d_m$, which we denote by q_{m+1}.

LEMMA 4.1. *Define* $q_{m+1} = A(x_m - d_m) - b$. *Then for* $m = 0, 1, 2, \ldots,$

$$|q_{m+1}| \leqslant uC_6|A||x_m - x| + (n\bar{u} + c_3u^2)|A||x| + \bar{u}uC_7|A||A^{-1}||A||x|,$$

assuming $c_1u\kappa(A^{-1}) \leqslant \frac{1}{2}$, *where*

$$\bar{u} = \begin{cases} u & \text{for s.p.r.a.,} \\ u^2 & \text{for d.p.r.a.} \end{cases}$$

Note. The quantities

$$c_3 = \begin{cases} (1+u)[(1+u)^n - 1]/u^2 - n/u & \text{for s.p.r.a.,} \\ (1+u)(1+u^2)[(1+u^2)^n - 1]/u^2 - n & \text{for d.p.r.a.,} \end{cases}$$

$$c_4 = \begin{cases} 0 & \text{for s.p.r.a.,} \\ 1+u & \text{for d.p.r.a.,} \end{cases}$$

$$c_5 = n + (c_3 + c_4)u^2/\bar{u},$$

$$C_6 = C_2 + (1 + c_5\bar{u}/u)(I + uC_2|A||A^{-1}|),$$

$$C_7 = (n + c_3u^2/\bar{u})C_2,$$

are bounded above by functions of n only.

Proof. The computed residual

$$r_m = Ax_m - b + f_m, \tag{4.1}$$

where

$$\begin{aligned} |f_m| &\leqslant (n\bar{u} + c_3u^2)|A||x_m| + (u + c_4u^2)|Ax_m - b| \\ &\leqslant (n\bar{u} + c_3u^2)|A||x| + c_5\bar{u}|A||x_m - x| + u|A(x_m - x)|. \end{aligned}$$

From (3.1) the computed error d_m satisfies

$$Ad_m - r_m = g_m, \quad \text{where } |g_m| \leqslant uC_2|A||A^{-1}r_m| \text{ if } c_1u\kappa(A^{-1}) \leqslant \tfrac{1}{2}.$$

Using (4.1) to eliminate r_m gives

$$A(x_m - d_m) - b = f_m - g_m, \quad \text{where } |g_m| \leqslant uC_2|A|(|x_m - x| + |A^{-1}||f_m|),$$

whence

$$|q_{m+1}| \leqslant uC_2|A||x_m - x| + (I + uC_2|A||A^{-1}|)|f_m|,$$

from which the theorem follows. □

In the next lemma we obtain bounds on the residual and error for x_{m+1} in terms of q_{m+1}.

LEMMA 4.2. *For* $m = 0, 1, 2, \ldots,$

$$|A(x_{m+1} - x)| \leqslant (I + u|A||A^{-1}|)|q_{m+1}| + u|A||x|$$

and

$$|x_{m+1} - x| \leqslant (1+u)|A^{-1}||q_{m+1}| + u|x|.$$

Proof. The new iterate $x_{m+1} = x_m - d_m + h_{m+1}$, where $|h_{m+1}| \leqslant u|x_m - d_m|$. Equivalently $x_{m+1} = A^{-1}q_{m+1} + x + h_{m+1}$, where $|h_{m+1}| \leqslant u|A^{-1}q_{m+1}| + u|x|$, from which the lemma follows. □

LEMMA 4.3. *For* $m = 0, 1, 2, \ldots,$

$$|q_{m+1}| \leqslant (uC_8|A||A^{-1}|)^m uC_{10}|A||x| + n\bar{u}|A||x| + u^2C_{11}|A||x| + \bar{u}uC_{12}|A||A^{-1}||A||x|,$$

assuming $c_8 u\kappa(A^{-1}) \leqslant \frac{1}{2}$.

Note. The quantities

$$\begin{aligned} C_8 &= (1+u)C_6, \\ c_8 &= \|C_8\|, \\ C_9 &= C_6 + c_3 I, \\ C_{10} &= C_6 + (n\bar{u}/u + c_3 u)I + \bar{u}C_7|A||A^{-1}|, \\ C_{11} &= (I - uC_8|A||A^{-1}|)^{-1}C_9, \\ C_{12} &= (I - uC_8|A||A^{-1}|)^{-1}(nC_8 + C_7), \end{aligned}$$

are bounded above by functions of n only.

Proof. Substituting the second inequality of Lemma 4.2 into that of Lemma 4.1 yields

$$|q_{m+1}| \leqslant uC_8|A||A^{-1}||q_m| + n\bar{u}|A||x| + u^2C_9|A||x| + \bar{u}uC_7|A||A^{-1}||A||x|,$$

assuming $c_1 u\kappa(A^{-1}) \leqslant \frac{1}{2}$. The proof is completed by induction on m. The lemma is true for $m = 0$ because of Lemma 4.1. Assume it is true for m. Then it is also true for $m+1$ because of the above bound on $|q_{m+1}|$ in terms of $|q_m|$. □

THEOREM 4.4. *For* $m = 0, 1, 2, \ldots,$

$$\begin{aligned} |A(x_{m+1} - x)| \leqslant\ & C_{13}(uC_8|A||A^{-1}|)^m uC_{10}|A||x| \\ & + (u + n\bar{u})|A||x| + u^2C_{14}|A||x| \\ & + \bar{u}uC_{15}|A||A^{-1}||A||x| \end{aligned}$$

and

$$\begin{aligned} |x_{m+1} - x| \leqslant\ & |A^{-1}|(uC_8|A||A^{-1}|)^m uC_{16}|A||x| + u|x| \\ & + n\bar{u}|A^{-1}||A||x| + u^2|A^{-1}|C_{17}|A||x| \\ & + \bar{u}u|A^{-1}|C_{18}|A||A^{-1}||A||x|, \end{aligned}$$

assuming $c_8 u\kappa(A^{-1}) \leqslant \frac{1}{2}$.

Note. The quantities

$$\begin{aligned} C_{13} &= I + u|A||A^{-1}|, \\ C_{14} &= C_{13}C_{11}, \\ C_{15} &= nI + C_{13}C_{15}, \\ C_{16} &= (1+u)C_{10}, \\ C_{17} &= (n\bar{u}/u)I + (1+u)C_{11}, \\ C_{18} &= (1+u)C_{12}, \end{aligned}$$

are bounded above by functions of n only.

Proof. Follows from substituting the inequality of Lemma 4.3 into those of Lemma 4.2. □

We note that iterating until convergence yields a qualitatively better error bound than performing a fixed number of iterations because for any fixed value of m the "iteration error" (or "convergence error") term cannot be absorbed into any of the other terms. This is true for *both* s.p.r.a. and d.p.r.a. Also, the error bounds indicate that convergence comes more quickly for s.p.r.a. than for d.p.r.a., which is not surprising.

From Section 3 and Theorem 4.4 we get the following bounds for the norm of the residual:

$$\|Ax_1 - b\| \leqslant c_2 u \| |A| |x| \|,$$

$$\overline{\lim_{m\to\infty}} \|Ax_m - b\| \leqslant (n+1)u \| |A| |x| \| + c_{19} u^2 \| |A| |A^{-1}| |A| |x| \| \quad \text{for s.p.r.a.,}$$

and

$$\overline{\lim_{m\to\infty}} \|Ax_m - b\| \leqslant (u + c_{20} u^2) \| |A| |x| \| \quad \text{for d.p.r.a.,}$$

provided that $c_8 u \kappa(A^{-1}) \leqslant \tfrac{1}{2}$. For the norm of the error we get the bounds

$$\|x_1 - x\| \leqslant c_2 u \|A^{-1}\| \| |A| |x| \|,$$

$$\overline{\lim_{m\to\infty}} \|x_m - x\| \leqslant nu \| |A^{-1}| |A| |x| \| + u\|x\| + c_{21} u^2 \|A^{-1}\| \| |A| |A^{-1}| |A| |x| \| \quad \text{for s.p.r.a.,}$$

and

$$\overline{\lim_{m\to\infty}} \|x_m - x\| \leqslant u\|x\| + c_{22} u^2 \|A^{-1}\| \| |A| |x| \| \quad \text{for d.p.r.a.,}$$

provided that $c_8 u \kappa(A^{-1}) \leqslant \tfrac{1}{2}$.

Note. The quantities

$$\begin{aligned}
c_2 &= \|C_2\|, \\
c_{19} &= \|C_{14} |A| |x| \| / \| |A| |A^{-1}| |A| |x| \| + \|C_{15}\|, \\
c_{20} &= \|nI + C_{14} + uC_{15} |A| |A^{-1}|\|, \\
c_{21} &= \|C_{17} |A| |x| \| / \| |A| |A^{-1}| |A| |x| \| + \|C_{18}\|, \\
c_{22} &= \|nI + C_{17} + uC_{18} |A| |A^{-1}|\|,
\end{aligned}$$

are bounded above by functions of n only.

Remark. From Theorems 4.1 and 3.1 of Jankowski and Woźniakowski (1977) one obtains the following results: Consider a linear equation solver which determines a solution $\hat{x}$ satisfying

$$\|\hat{x} - x\| \leqslant c_{23} u \operatorname{cond}(A) \|x\|,$$

assuming that the $c_{23} u \operatorname{cond}(A) \leqslant \tfrac{1}{2}$, where $\operatorname{cond}(A) = \|A^{-1}\| \|A\|$. The use of iterative refinement with this linear equation solver yields iterates x_{m+1}, $m = 1, 2, \ldots,$

for which the residual norm

$$\|Ax_{m+1} - b\| \leqslant \begin{cases} (c_{24}u + c_{25}u^2 \operatorname{cond}^2(A))\|A\|\|x\| & \text{for s.p.r.a.,} \\ c_{26}u\|A\|\|x\| & \text{for d.p.r.a.} \end{cases}$$

and the error norm

$$\|x_{m+1} - x\| \leqslant u\|x\| + c_{27}\bar{u} \operatorname{cond}(A)\|x\| + c_{28}(c_{29}u \operatorname{cond}(A))^{m+1}\|x\|$$

provided that $c_{29}u \operatorname{cond}(A) \leqslant \frac{1}{2}$. The principal way in which these bounds differ from the ones following Theorem 4.4 is the use of norms in place of absolute values. Another significant difference is the term $c_{25}u^2\operatorname{cond}^2(A)$ instead of $c_{19}u^2\kappa(A^{-1})$ in the bound on the residual norm, which is due to different assumptions on the accuracy of the linear equation solver.

5. Backward Error Bounds. We establish the numerical stability of Gaussian elimination with iterative refinement by suitably bounding the backward error η. Similar bounds on η^R and η^E can be obtained by substituting the bounds of Theorem 4.4 into those of Theorems 2.2 and 2.3. The best possible bounds of the type we are seeking lead to monstrous expressions that are difficult to interpret. Thus, we compromise by using only the quantity $\kappa(A^{-1})$ and the quantity

$$\sigma(A, x) = \max \frac{e\||A||x|\|}{\|A\|x|\|},$$

which was introduced at the end of Section 3 as a measure of bad scaling for column pivoting. For example, the quantity

$$\max \frac{\|A\|\|A^{-1}\|\|A\|x|\|}{\|A\|x|\|}$$

is replaced by $\kappa(A^{-1})\sigma(A, x)$ even though this may be a severe overestimate.

THEOREM 5.1. *For* $m = 0, 1, 2, \ldots,$

$$\eta_{m+1} \leqslant c_{32}u(c_8u(A^{-1}))^m\sigma(A, x) + u + n\bar{u} + c_{33}u^2\sigma(A, x) + c_{34}\bar{u}u\kappa(A^{-1})\sigma(A, x),$$

assuming that

$$c_{16}u(c_8u\kappa(A^{-1}))^m\kappa(A^{-1})\sigma(A, x) + c_8u\kappa(A^{-1}) + c_{31}\bar{u}\kappa(A^{-1})\sigma(A, x) \leqslant \tfrac{1}{2}.$$

Note. The quantities

$$\begin{aligned} c_{30} &= \|C_{13}\|\|C_{10}\|, \\ c_{14} &= \|C_{14}\|, \\ c_{15} &= \|C_{15}\|, \\ c_{16} &= \|C_{16}\|, \\ c_{31} &= n + (u/\bar{u})\|C_{17}\| + u\|C_{18}\|\kappa(A^{-1}), \\ c_{32} &= 2[c_{30} + (u + n\bar{u})c_{16}\kappa(A^{-1})], \end{aligned}$$

$$c_{33} = 2[c_{14} + (u + n\bar{u})/(u\sigma(A, x))],$$
$$c_{34} = 2[c_{15} + (u + n\bar{u})/u],$$

are bounded above by functions of n only.

Proof. It follows from Theorem 2.2 that

$$\eta_{m+1} \leqslant \left(\max \frac{|A(x_{m+1} - x)|}{|A||x|}\right) \Big/ \left(1 - \max \frac{|A||x_{m+1} - x|}{|A||x|}\right).$$

Theorem 4.4 yields the bound

$$\max \frac{|A(x_{m+1} - x)|}{|A||x|} \leqslant c_{30}u(c_8 u\kappa(A^{-1}))^m \sigma(A, x) + u + n\bar{u} + c_{14}u^2\sigma(A, x) + c_{15}\bar{u}u\kappa(A^{-1})\sigma(A, x)$$

and the bound

$$\max \frac{|A||x_{m+1} - x|}{|A||x|} \leqslant c_{16}u(c_8 u\kappa(A^{-1}))^m \kappa(A^{-1})\sigma(A, x) + u + c_{31}\bar{u}\kappa(A^{-1})\sigma(A, x),$$

from which the theorem follows. □

Without any iterations of iterative refinement we have

$$\eta_1 \leqslant c_{35}u\sigma(A, x) \quad \text{if } c_{36}\kappa(A^{-1})\sigma(A, x) \leqslant \tfrac{1}{2},$$

but for one iteration with s.p.r.a.

$$\eta_2 \leqslant (n + 1)u + c_{37}u^2\kappa(A^{-1})\sigma(A, x) \quad \text{if } c_{36}u\kappa(A^{-1})\sigma(A, x) \leqslant \tfrac{1}{2}.$$

Hence, just one iteration of iterative refinement with just single precision accumulation of the residuals is enough to make Gaussian elimination stable. This may seem to contradict the usual advice [Forsythe and Moler (1967), p. 49] that "It is absolutely essential that this residual r_k be computed with a higher precision than that of the rest of the computation." Actually, there is little conflict because it has been shown that poorly scaled systems may be solved with an effective precision of much less than single precision. However, the restriction $cu\kappa(A^{-1})\sigma(A, x) \leqslant \tfrac{1}{2}$ indicates that a big reduction in the backward error may not be realized for badly scaled problems which are very ill-conditioned unless the precision is high enough.

Note. The quantities

$$c_{35} = c_{32} + 1 + n\bar{u}/u + c_{33}u + c_{34}\bar{u}\kappa(A^{-1}),$$
$$c_{36} = c_{16} + c_8/\sigma(A, x) + c_{31}\bar{u}/u,$$
$$c_{37} = c_{32}c_8 + c_{33}/\kappa(A^{-1}) + c_{34}\bar{u}/u,$$

are bounded above by functions of n only.

For iteration until convergence with d.p.r.a.

$$\overline{\lim_{m\to\infty}}\, \eta_m \leqslant u + c_{38}u^2\sigma(A, x) \quad \text{if } c_8 u\kappa(A^{-1}) + c_{39}u\sqrt{\kappa(A^{-1})\sigma(A, x)} \leqslant \tfrac{1}{2}.$$

This is a significant improvement over the single precision case due to the relaxation of the restriction on the size of $\kappa(A^{-1})$ and $\sigma(A, x)$. This is an important advantage for problems which are both poorly scaled and ill-conditioned.

Note. The quantities

$$c_{38} = n/\sigma(A, x) + c_{33} + c_{34}u\kappa(A^{-1}),$$
$$c_{39} = \sqrt{c_{31}/2},$$

are bounded above by functions of n only.

6. Gaussian Elimination with Row Pivoting. Sections 3, 4, and 5 consider the use of column pivoting; this section considers instead row pivoting and complete pivoting.

Here the basic result, shown in Skeel (1979), upon which our error analysis depends, is that the solution $\hat{x}$ computed by row pivoting satisfies

$$|A\hat{x} - b| \leqslant u|A|C_{51}|\hat{x}|.$$

Straightforward modifications of the results for column pivoting yield corresponding results for row pivoting. For example, Theorem 4.4 holds for row pivoting if $|A|C_{50+j}$ is substituted for every occurrence of $C_j|A|$ and $\kappa(A)$ for $\kappa(A^{-1})$. Without iterative refinement we obtain the following stability bounds:

$$\eta \leqslant c_{51}u \max \frac{|A||e|\|x\|}{|A||x|} + O(u^2),$$

$$\eta^R \leqslant c_{51}u \frac{\|A\|\|x\|}{\||A||x|\|} + O(u^2),$$

$$\eta^E \leqslant c_{51}u \frac{\||A^{-1}||A|\|\|x\|}{\||A^{-1}||A||x|\|} + O(u^2).$$

What is most interesting about these bounds is that it seems that row pivoting is not R-stable. To see that this is actually true, consider

$$A = \begin{bmatrix} 3 & 3 \\ 3 \times 10^N & 0 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Using rounded t-digit decimal arithmetic, the computed solution is

$$\hat{x} = \begin{bmatrix} .33\cdots3 \times 10^{-t} \\ .33\cdots3 \end{bmatrix}.$$

For the maximum norm the backward error

$$\eta^R = \frac{\|A\hat{x} - b\|_\infty}{\||A||x|\|_\infty} + O(10^{-2t}) = 10^N \times 10^{-t} + O(10^{-2t}),$$

which is an arbitrarily large multiple of the unit roundoff error $u = \frac{1}{2} \times 10^{1-t}$.

Concerning stability or E-stability, either column or row pivoting could be arbitrarily better than the other. Nevertheless, it is interesting that the error bounds for row pivoting contain the quantity $\kappa(A)$, which also arises in the bounds for η^E given

in Theorem 2.3; whereas, the error bounds for column pivoting contain $\kappa(A^{-1})$, which can be arbitrarily different from $\kappa(A)$. For example, for

$$A = \begin{bmatrix} 1 & M \\ 1 & M+1 \end{bmatrix},$$

$\kappa_\infty(A^{-1}) = 4M + 3$ while $\kappa_\infty(A) = 2M^2 + 4M + 1$.

Row pivoting (like column pivoting) can be made stable by an appropriate scaling of the columns; but because the proper scaling cannot be determined efficiently, iterative refinement could be useful for eliminating the effects of the poor scaling.

Complete pivoting is both column and row pivoting, and it satisfies the error bounds of both. Hence, the convergence of iterative refinement for complete pivoting requires only that either $c_8 u\kappa(A^{-1}) \leqslant \frac{1}{2}$ or $c_{58} u\kappa(A) \leqslant \frac{1}{2}$. The example just given shows that this requirement is much less restrictive than the convergence condition for either type of partial pivoting.

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

1. F. L. BAUER, "Optimally scaled matrices," *Numer. Math.*, v. 5, 1963, pp. 78–87.

2. F. L. BAUER, J. STOER & C. WITZGALL, "Absolute and monotonic norms," *Numer. Math.*, v. 3, 1961, pp. 257–264.

3. G. E. FORSYTHE, M. A. MALCOLM & C. B. MOLER, *Computer Methods for Mathematical Computations*, Prentice-Hall, Englewood Cliffs, N. J., 1977.

4. G. E. FORSYTHE & C. B. MOLER, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, N. J., 1967.

5. C. W. GEAR, *Numerical Errors in Sparse Linear Equations*, File F-75-885, Dept. of Computer Sci., Univ. of Illinois at Urbana-Champaign, May 1975.

6. R. W. HAMMING, *Introduction to Applied Numerical Analysis*, McGraw-Hill, New York, 1971.

7. M. JANKOWSKI & M. WOŹNIAKOWSKI, "Iterative refinement implies numerical stability," *BIT*, v. 17, 1977, pp. 303–311.

8. W. MILLER, "On the stability of finite numerical procedures," *Numer. Math.*, v. 19, 1972, pp. 425–432.

9. W. MILLER, Private communication, 1977.

10. W. MILLER & C. WRATHALL, *Software for Roundoff Analysis of Matrix Algorithms*, Academic Press, New York. (To appear.)

11. W. OETTLI & W. PRAGER, "Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides," *Numer. Math.*, v. 6, 1964, pp. 405–409.

12. R. D. SKEEL, *Gaussian Elimination and Numerical Instability*, Report R-77-862, Dept. of Computer Sci., Univ. of Illinois at Urbana-Champaign, April 1977.

13. R. D. SKEEL, "Scaling for numerical stability in Gaussian elimination," *J. Assoc. Comput. Mach.*, v. 26, 1979, pp. 494–526.

14. G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

15. J. H. WILKINSON, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, N. J., 1963.