How Accurate is Gaussian Elimination?*

Nicholas J. Higham

TR 89-1024
July 1989

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501

*To appear in the proceedings of the 13th Dundee Biennial Conference on Numerical Analysis, D.F. Griffiths and G.A. Watson, eds., Longman, 1990. Appears also as a University of Manchester Numerical Analysis report.

How Accurate is Gaussian Elimination?*

Nicholas J. Higham†
Department of Computer Science
Cornell University
Ithaca, New York 14853

Abstract

J.H. Wilkinson put Gaussian elimination (GE) on a sound numerical footing in the 1960s when he showed that with partial pivoting the method is stable in the sense of yielding a small backward error. He also derived bounds proportional to the condition number $\kappa(A)$ for the forward error $\|x - \widehat{x}\|$, where $\widehat{x}$ is the computed solution to $Ax = b$. More recent work has furthered our understanding of GE, largely through the use of componentwise rather than normwise analysis. We survey what is known about the accuracy of GE in both the forward and backward error senses. Particular topics include: classes of matrix for which it is advantageous not to pivot; how to estimate or compute the backward error; iterative refinement in single precision; and how to compute efficiently a bound on the forward error.

Key words: Gaussian elimination, partial pivoting, rounding error analysis, backward error, forward error, condition number, iterative refinement in single precision, growth factor, componentwise bounds, condition estimator.

AMS(MOS) Subject classifications. primary 65F05, 65G05.

1 Introduction

Consider the linear system $Ax = b$, where A is the 7×7 Vandermonde matrix with $a_{ij} = j^{i-1}$ and $b_i = i$. Using Matlab we solved this system in single precision (unit roundoff $\approx 10^{-7}$) via both Gaussian elimination (GE) and Gaussian elimination with

*To appear in the proceedings of the 13th Dundee Biennial Conference on Numerical Analysis, D.F. Griffiths and G.A. Watson, eds., Longman, 1990.

†On leave from the University of Manchester.

partial pivoting (GEPP), obtaining computed solutions $\widehat{x}_{\mathrm{GE}}$ and $\widehat{x}_{\mathrm{GEPP}}$, respectively. We found that

$$\frac{\|x - \widehat{x}_{\mathrm{GE}}\|_\infty}{\|x\|_\infty} \approx 4 \times 10^{-8},$$
$$\frac{\|x - \widehat{x}_{\mathrm{GEPP}}\|_\infty}{\|x\|_\infty} \approx 6 \times 10^{-3}.$$

An alternative measure of the quality of each of these computed solutions is whether it is the exact solution of a perturbed system $(A + \Delta A)x = b$ for some small ΔA (clearly, ΔA is not unique). Using results to be described in section 2 we found that the minimum value of $\|\Delta A\|_\infty / \|A\|_\infty$ is approximately 3×10^{-9} for both $\widehat{x}_{\mathrm{GE}}$ and $\widehat{x}_{\mathrm{GEPP}}$. However, the minimum value of $\max_{i,j} |\Delta a_{ij}/a_{ij}|$, which measures the perturbation componentwise relative to A, is 3×10^{-8} for $\widehat{x}_{\mathrm{GE}}$ and 2×10^{-6} for $\widehat{x}_{\mathrm{GEPP}}$. If we do one step of iterative refinement starting from $\widehat{x}_{\mathrm{GEPP}}$, *entirely in single precision*, we obtain an updated solution $\overline{x}$ for which the componentwise measure of the size of ΔA is 5×10^{-8} and $\|x - \overline{x}\|_\infty / \|x\|_\infty \approx 4 \times 10^{-5}$.

This example opposes the conventional wisdom that GE with partial pivoting is preferable to GE without pivoting. It also shows that iterative refinement in single precision can be beneficial. It is natural to ask: Can this behaviour be explained by a priori analysis of the problem, and can the example be generalized? More generally, what can one say about the sizes of $x - \widehat{x}$ and ΔA for arbitrary A and b?

Answers to these questions are contained in this survey of the accuracy of Gaussian elimination in finite precision arithmetic. Work on this subject began in the 1940s at around the time of the first electronic computers. It reached maturity in the 1960s, largely due to Wilkinson's contributions. Research done since the mid 1970s has provided further understanding of the subject.

We begin in section 2 by discussing in detail the notion of backward error and showing how a wide class of backward errors can be computed. In section 3 we survey rounding error analysis for Gaussian elimination. To show that error analysis need not be difficult we give a derivation of one of the most useful backward error bounds. Implications of the analysis are discussed in section 4, while section 5 covers practical computation of backward error bounds. Iterative refinement, which has attracted renewed interest recently, is the subject of section 6. In section 7 we turn our attention to estimation of the forward error. We offer some final thoughts in section 8, where we show how the results presented here help to explain the numerical example above.

Our notation is as follows. Throughout, A is a real $n \times n$ matrix. Computed quantities are denoted with a hat. Thus $\widehat{x}$ is a computed solution to $Ax = b$ and $\widehat{L}$, $\widehat{U}$ are computed LU factors of A.

Much of this paper is concerned with backward error analysis. This is because there is a lot to say about the topic, not because we believe backward error analysis is of vital importance to all users of GE software. We suspect that many users are interested mainly in the size of $x - \widehat{x}$. For them the most important material is in section 7, and familiarity with sections 2 and 6 is recommended.

2 Backward Error

In analysing the behaviour of GE in finite precision arithmetic two types of error are of interest. The *forward error* is any suitable norm of $x - \widehat{x}$, and is easy to appreciate. *Backward error* is a more subtle concept, as we now explain.

In GE backward error is a measure of the smallest perturbations such that either

$$\widehat{L}\widehat{U} = A + \Delta A \tag{2.1}$$

if we are concerned with the LU factorization alone, or

$$(A + \Delta A)\widehat{x} = b + \Delta b \tag{2.2}$$

for the overall process. In (2.1) ΔA is unique and the question is how to measure its size. In (2.2) there are many possible ΔA and Δb, which makes determination of the backward error nontrivial.

The usual motivation for considering backward error is that there may already be uncertainty in the data A and b (arising from rounding errors in storing or computing the data, for example). If the backward error and the uncertainty are of similar magnitude then it can be argued that the computed solution $\widehat{x}$ is beyond reasonable criticism. Another attractive feature of backward error analysis is that it enables one to invoke existing perturbation theory to produce bounds for the forward error (see section 7).

We consider two ways to measure the size of the perturbations ΔA and Δb.

(a) Normwise Backward Error

Here we measure the perturbations using norms. We will use an arbitrary vector norm and the corresponding subordinate matrix norm. For the LU factorization the backward error is simply $\beta_N = \|A - \widehat{L}\widehat{U}\| / \|A\|$. For $\widehat{x}$ it is

$$\begin{aligned}\beta_N \;=\; \min\{\omega : (A + \Delta A)\widehat{x} = b + \Delta b,\ & \|\Delta A\| \le \omega\|A\|,\ \|\Delta b\| \le \omega\|b\|, \\ & \Delta A \in \mathbb{R}^{n \times n},\ \Delta b \in \mathbb{R}^n\}.\end{aligned}$$

The following result of Rigal and Gaches [34] gives a convenient expression for β_N.

Theorem 2.1

$$\beta_N = \frac{\|r\|}{\|A\|\,\|\widehat{x}\| + \|b\|}, \tag{2.3}$$

and perturbations which achieve the minimum in the definition of β_N are

$$\begin{aligned}\Delta A_{\min} &= \frac{\|A\|\|\widehat{x}\|}{\|A\|\,\|\widehat{x}\|+\|b\|} r z^T,\\ \Delta b_{\min} &= \frac{\|b\|}{\|A\|\,\|\widehat{x}\|+\|b\|} r,\end{aligned}$$

where $r = b - A\widehat{x}$ and z is a vector dual to $\widehat{x}$, that is,

$$z^T\widehat{x} = \|z\|_D\|\widehat{x}\| = 1 \qquad \textit{where} \qquad \|z\|_D = \max_{y\neq 0}\frac{|z^T y|}{\|y\|}.$$

Proof. See [34, Theorem 1]. ∎

If we set $\Delta b = 0$ in the definition of β_N then the result simplifies to $\beta_N = \|r\|/(\|A\|\|\widehat{x}\|)$ and $\Delta A_{\min} = rz^T$. In the case of the 2-norm, $z = \widehat{x}/\|\widehat{x}\|_2^2$ and the result with $\Delta b \equiv 0$ is well-known.

If A has some special property, then in the definition of β_N we might wish to prescribe that $A + \Delta A$ has the same property. In the case of symmetry, this more restrictive backward error measure (with $\Delta b \equiv 0$) has been investigated by Bunch, Demmel and Van Loan [4]. They show the pleasing result that the symmetry constraint does not change the backward error for the 2-norm, and it increases it by at most a factor $\sqrt{2}$ for the Frobenius norm.

(b) Componentwise Backward Error

This definition involves a matrix $E \geq 0$ and a vector $f \geq 0$ whose elements provide relative tolerances against which the components of ΔA and Δb are measured. For $\widehat{x}$, the componentwise backward error is defined as

$$\begin{aligned}\beta_C(E,f) = \min\{\omega : (A+\Delta A)\widehat{x} = b + \Delta b,\ |\Delta A| \leq \omega E,\ |\Delta b| \leq \omega f,\\ \Delta A \in \mathbb{R}^{n\times n},\ \Delta b \in \mathbb{R}^n\}.\end{aligned}$$

The matrix absolute value and matrix inequality are interpreted componentwise: thus $|X| \leq Y$ means that $|x_{ij}| \leq y_{ij}$ for all i,j. For the LU factorization the definition is simply $\beta_C(E) = \max\{|\Delta a_{ij}|/e_{ij},\ 1 \leq i,j \leq n\}$. Two extreme choices of E and f are as follows.

1. $E = |A|$, $f = |b|$. In this case we measure the size of the perturbation in each element of A or b relative to the element itself. This is the most stringent backward error measure of general interest. Note that the constraints in the definition of $\beta_C(|A|,|b|)$ force $A + \Delta A$ and $b + \Delta b$ to have the same sparsity patterns as A and b respectively. Following [1] we call $\beta_C(|A|,|b|)$ the *componentwise relative backward error.*

2. $E = \|A\|_\infty ee^T$, $f = \|b\|_\infty e$, where $e = (1, 1, \ldots, 1)^T$. For this choice we are measuring perturbations in an absolute sense (in a similar, but not identical, way to the norm case (a)).

An expression for β_C is given by Oettli and Prager [29], and we present the short proof since it aids in understanding the result.

Theorem 2.2

$$\beta_C(E, f) = \max_i \frac{|b - A\hat{x}|_i}{(E|\hat{x}| + f)_i}, \tag{2.4}$$

where $0/0$ *is interpreted as zero, and* $\xi/0$ $(\xi \neq 0)$ *as infinity, the latter case meaning that no finite* ω *exists in the definition of* $\beta_C(E, f)$.

Proof. For any candidate perturbations ΔA and Δb in the definition of $\beta_C(E, f)$ we have

$$|r| = |b - A\hat{x}| = |\Delta A\hat{x} - \Delta b| \leq \omega E|\hat{x}| + \omega f,$$

which implies that

$$\beta_C(E, f) \geq \max_i \frac{|r_i|}{(E|\hat{x}| + f)_i} \equiv \theta.$$

To show that this lower bound is attained note that

$$r = D(E|\hat{x}| + f) \tag{2.5}$$

for a diagonal D with $|D| \leq \theta I$. Defining $\Delta A = DE\,\mathrm{diag}(\mathrm{sign}(\hat{x}))$ and $\Delta b = -Df$ we have $|\Delta A| \leq \theta E$, $|\Delta b| \leq \theta f$, and, using (2.5),

$$(A + \Delta A)\hat{x} - (b + \Delta b) = A\hat{x} + DE|\hat{x}| - b + Df = 0,$$

as required. ∎

This result shows that β_C can be computed using two matrix-vector multiplies. For the E and f in case (2) above the formula reduces to $\beta_C = \|r\|_\infty / (\|A\|_\infty \|\hat{x}\|_1 + \|b\|_\infty)$, which is very similar to β_N for the ∞-norm.

3 Error Analysis of Gaussian Elimination

In this section we give a brief survey of rounding error analysis for Gaussian elimination. For further perspective on this topic we recommend the papers [31, 51, 52, 53] of Wilkinson.

Unless otherwise stated, results quoted are for floating point arithmetic. We will ignore permutations in stating backward error bounds; thus A actually denotes the original matrix after all row or column interchanges necessary for the pivoting strategy have been performed.

In the 1940s there were three major papers giving error analyses of GE. Hotelling [27] presented a short forward error analysis of the LU factorization stage of GE. Under the assumptions that $|a_{ij}| \le 1$ and $|b_i| \le 1$ for all i and j, and that the pivots are all of modulus unity, Hotelling derives a bound containing a factor 4^{n-1} for the error in the elements of the reduced upper triangular system. This result led to pessimism about the practical effectiveness of GE for solving large systems of equations. Three papers later in the same decade helped to restore confidence in GE.

Von Neumann and Goldstine [46] gave a long and difficult fixed-point error analysis for the inversion of a symmetric positive definite matrix A via GE. They obtained a bound proportional to $\kappa_2(A)$ for the residual $\|A\widehat{X} - I\|_2$ of the computed inverse $\widehat{X}$. Wilkinson [51] gives an interesting critique of this paper and points out that the residual bound could hardly be improved using modern error analysis techniques.

Turing [44] analysed GEPP for general matrices and obtained a bound for $\|x - \widehat{x}\|_\infty$ that contains a term proportional to $\|A^{-1}\|_\infty^2$. (By making a trivial change in the analysis Turing's bound can be made proportional only to $\|A^{-1}\|_\infty$.) Turing also showed that the factor 4^{n-1} in Hotelling's bound can be improved to 2^{n-1} and that still the bound is attained only in exceptional cases.

Fox, Huskey and Wilkinson [18] presented empirical evidence in support of GE, commenting that "in our practical experience on matrices of orders up to the twentieth, some of them very ill-conditioned, the errors were in fact quite small".

A major breakthrough in the error analysis of GE came with Wilkinson's pioneering backward error analysis [48, 49]. Wilkinson showed that with partial or complete pivoting the computed solution $\widehat{x}$ satisfies

$$(A + E)\widehat{x} = b, \tag{3.1}$$

where

$$\|E\|_\infty \le \rho_n p(n) u \|A\|_\infty. \tag{3.2}$$

Here, p is a cubic polynomial and the growth factor ρ_n is defined by

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

where the $a_{ij}^{(k)}$, $k = 0, 1, \ldots, n-1$, are the elements that occur during the elimination. Apart from its simplicity and elegance, the main feature that distinguishes Wilkinson's analysis from the earlier error analyses of GE is that it bounds the normwise backward error rather than the forward error or the residual.

Three of the first textbooks to incorporate Wilkinson's analysis were those of Fox [17, pp. 161–174], Wendroff [47] and Forsythe and Moler [16, Ch. 21]. Fox gives a simplified analysis for fixed-point arithmetic under the assumption that the growth factor is of order 1. Forsythe and Moler give a particularly readable backward error analysis which has been widely quoted.

Wilkinson's 1961 result is essentially the best that can be obtained by a normwise analysis. Subsequent work in error analysis for GE has mainly been concerned with bounding the backward error componentwise, that is, obtaining n^2 individual bounds for the elements of the backward error matrix E, rather than a single bound for a norm of E.

Chartres and Geuder [6] analyse the Doolittle "compact" version of GE. They derive the componentwise backward error result

$$(A+E)\hat{x} = b,$$
$$|e_{ij}| \le 2(j+3)w_{ij}c_1(n,u) + \begin{cases} 2|\hat{u}_{ij}|c_2(n,u), & i<j, \\ 3|\hat{u}_{jj}|c_2(n,u), & i=j, \\ 2|\hat{l}_{ij}\hat{u}_{jj}|c_2(n,u), & i>j, \end{cases} \tag{3.3}$$

where $c_1(n,u) \approx u$, $c_2(n,u) \approx (n+1)u$, and

$$w_{ij} = \sum_{k=1}^{m-1} |\hat{l}_{ik}\hat{u}_{kj}| = (|\hat{L}||\hat{U}|)_{ij} - |\hat{l}_{im}\hat{u}_{mj}|, \qquad m = \min(i,j).$$

We note that Wilkinson could have given a componentwise bound in place of (3.2), since most of his analysis is at the element level.

Reid [32] shows that the assumption in Wilkinson's analysis that partial pivoting or complete pivoting is used is unnecessary. Without making any assumptions on the pivoting strategy he derives the result for the LU factorization

$$\hat{L}\hat{U} = A + E,$$
$$|e_{ij}| \le 3.01n\, u \max_k |a_{ij}^{(k)}|.$$

Again, this is a componentwise bound. Note that the backward error analyses discussed so far display three different "styles" of error bound, indicating the considerable freedom one has in deciding how to develop and phrase a backward error analysis.

de Boor and Pinkus [9] give the result

$$(A+E)\hat{x} = b, \tag{3.4}$$
$$|E| \le \gamma_n(2+\gamma_n)|\hat{L}||\hat{U}|, \tag{3.5}$$

where

$$\gamma_n = \frac{nu}{1-nu}.$$

They refer to the original 1972 German edition of [40] for a proof of the result, and explain several advantages to be gained by working with the matrix-level bound (3.5) (see section 4.2).

We briefly mention some other relevant work:

- Demmel [10] shows how existing backward error analyses for GE can be modified to take into account the possibility of underflow (the analyses we have described assume that underflow does not occur).

- Backward error analysis has been used to investigate how the accuracy and stability of GE are affected by scaling, and by use of row pivoting instead of column pivoting. van der Sluis [45] and Stewart [39] employ norm analysis, while Skeel [35, 37] uses a componentwise approach.

- Specialized backward error analyses have been done for the Cholesky factorization of positive (semi-) definite matrices; see [25] and the references therein.

- Forward error analyses have been done for GE. The analyses are more complicated and more difficult to interpret than the backward error analyses. See [30] and [41, 42].

The "$|\widehat{L}||\widehat{U}|$" componentwise-wise style of backward error analysis is now well-known, as evidenced by its presence in several textbooks [8, 19, 40]. To emphasize the simplicity of the analysis we give a short proof of (3.5). This proof is modelled on one in [40]. See also [7] for a similar presentation.

We will use the standard model of floating point arithmetic:

$$fl(x \operatorname{op} y) = (x \operatorname{op} y)(1+\delta), \qquad |\delta| \leq u, \quad \operatorname{op} = +, -, *, /, \tag{3.6}$$

where u is the unit roundoff. The low-level technical details can be confined to two lemmas.

Lemma 3.1 *If $|\delta_i| \leq u$ and $p_i = \pm 1$ for $1 \leq i \leq n$, and $nu < 1$, then*

$$\prod_{i=1}^{n} (1+\delta_i)^{p_i} = 1 + \theta_n,$$

where $|\theta_n| \leq \gamma_n \equiv nu/(1-nu)$.

Proof. A straightforward induction. ∎

Lemma 3.2 *If the expression $s = (c - \sum_{i=1}^{k-1} a_i b_i)/b_k$ is evaluated in floating point arithmetic, in whatever order, then the computed value $\widehat{s}$ satisfies*

$$|c - \sum_{i=1}^{k-1} a_i b_i - \widehat{s} b_k| \leq \gamma_k (|\widehat{s} b_k| + \sum_{i=1}^{k-1} |a_i||b_i|). \tag{3.7}$$

Proof. Straightforward manipulation. If we assume that the expression is evaluated in the natural left-right order then an intermediate result in the proof (and a backward error result in its own right) is

$$\hat{s}b_k(1+\theta_k) = c - \sum_{i=1}^{k-1} a_i b_i (1+\theta_i) \tag{3.8}$$

where the θ_i satisfy $|\theta_i| \le \gamma_i$. It is not hard to see that this inequality holds whatever ordering is used when evaluating s if we replace each θ_i on the right-hand side by θ_k (different in each instance). ∎

Recall that the GE algorithm comprises three nested loops, and that there are six ways of ordering the loops. Each version sustains precisely the same rounding errors under the model (3.6) because each version does the same floating point operations with the same arguments—only the order in which the operations are done differs. We find it convenient to analyse the Doolittle, or "*jik*" variant, which computes L a column at a time and U a row at a time according to

$$\begin{aligned} u_{ij} &= a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}, \qquad j \ge i \\ l_{ij} &= (a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj})/u_{jj}, \qquad i > j \end{aligned}$$

Applying Lemma 3.2 to these equations we obtain

$$\begin{aligned} |a_{ij} - \sum_{k=1}^{i} \hat{l}_{ik}\hat{u}_{kj}| &\le \gamma_i \sum_{k=1}^{i} |\hat{l}_{ik}\hat{u}_{kj}|, \qquad j \ge i, \\ |a_{ij} - \sum_{k=1}^{j} \hat{l}_{ik}\hat{u}_{kj}| &\le \gamma_j \sum_{k=1}^{j} |\hat{l}_{ik}\hat{u}_{kj}|, \qquad i > j, \end{aligned}$$

where we have defined $\hat{l}_{ii} = l_{ii} \equiv 1$. These inequalities may be written in matrix form as

$$\hat{L}\hat{U} = A + E, \qquad |E| \le \gamma_n |\hat{L}||\hat{U}|. \tag{3.9}$$

Finally, we have to consider the forward and back substitutions, $Ly = b$, $Ux = y$. From (3.8) we immediately obtain

$$(\hat{L} + \Delta\hat{L})\hat{y} = b, \qquad \Delta\hat{L} \le \gamma_n |\hat{L}|, \tag{3.10}$$

$$(\hat{U} + \Delta\hat{U})\hat{x} = \hat{y}, \qquad \Delta\hat{U} \le \gamma_n |\hat{U}|. \tag{3.11}$$

Thus $(\hat{L} + \Delta\hat{L})(\hat{U} + \Delta\hat{U})\hat{x} = b$, and combining (3.9)–(3.11) we arrive at (3.5). (Actually, we obtain (3.5) with the '2' replaced by '3'—a slightly more refined analysis of the substitution stages produces the '2'.)

Note that if A has bandwidth k ($a_{ij} = 0$ for $|i - j| > k$) then, since L and U have the same bandwidth, we can replace γ_n by γ_{k+1} in the above analysis.

4 Interpreting the Error Analysis

In this section we interpret the backward error analyses summarized in section 3 and look at their implications. It is instructive to make comparisons with the "ideal" bounds

$$\|E\|_\infty \le u\|A\|_\infty \qquad \text{(small normwise backward error)}, \tag{4.1}$$

$$|E| \le u|A| \qquad \text{(small componentwise relative backward error)}, \tag{4.2}$$

which hold if, for example, $A + E$ is the rounded version of A. (Note that (4.2) implies (4.1)).

4.1 Normwise Analysis

Wilkinson's backward error result is usually explained as follows. The bound (3.2) differs from the ideal bound (4.1) by having the extra factor $\rho_n p(n)$. The $p(n)$ term is fixed, and hence is beyond our control. Also, it is "pessimistic", since it arises from repeated use of triangle inequalities and taking of matrix norms. Wilkinson [49, pp. 62, 108] comments that the bound is usually not sharp even if we replace $p(n)$ by its square root. Therefore our attention is focussed on the size of the growth factor.

Let ρ_n, ρ_n^p, ρ_n^c denote the growth factors for GE with, respectively, no pivoting, partial pivoting and complete pivoting. It is easy to see that $\rho_n(A)$ can be arbitrarily large, and so GE without pivoting is unstable in general. For partial pivoting, ρ_n^p is almost invariably small in practice ($\rho_n^p < 10$, say) but a parametrized family of matrices is known for which it achieves its maximum of 2^{n-1} [26]. The situation is similar for ρ_n^c, except that a much smaller upper bound is known, and it has been conjectured that $\rho_n^c(A) \le n$ for real A. Recent work has shed more light on the behaviour of ρ_n^p and ρ_n^c. Higham and Higham [26] present several families of real matrices from practical applications for which $\rho_n^p(A)$ and $\rho_n^c(A)$ are about $n/2$. These examples show that moderately large growth factors can be achieved on non-contrived matrices. Trefethen and Schreiber [43] develop a statistical model of the average growth factor for partial pivoting and complete pivoting. Their model supports their empirical findings that for various distributions of random matrices the average growth factor (normalized by the standard deviation of the initial matrix elements) is close to $n^{2/3}$ for partial pivoting and $n^{1/2}$ for complete pivoting.

For certain classes of matrix special bounds are known for the growth factor (see [48], [50, pp. 218–220] and [38, p. 158]):

- If A is diagonally dominant by columns ($|a_{jj}| > \sum_{i \ne j} |a_{ij}|$ for all j) then $\rho_n(A) = \rho_n^p(A) \le 2$ (and no row interchanges are performed with partial pivoting).

- If A is tridiagonal then $\rho_n^p(A) \leq 2$ and if A is upper Hessenberg then $\rho_n^p(A) \leq n$.
- If A is symmetric positive definite then $\rho_n(A) \leq 1$.
- If the LU factors of A have nonnegative elements then $\rho_n(A) \leq 1$. $L \geq 0$ and $U \geq 0$ is guaranteed if A is *totally nonnegative*, that is, if the determinant of every submatrix of A is nonnegative.

4.2 Componentwise Analysis

The componentwise bound (3.5) is weaker than the ideal bound (4.2) in general, but it matches it (up to a factor n) if $|\widehat{L}||\widehat{U}| \leq c|A|$ with c a small constant, which is true in several cases as we now explain. We assume the use of GE *without* pivoting.

First, we mention the important case of triangular systems: here the componentwise relative backward error is always small, as shown by (3.10)–(3.11). The implications of this fact are explored in detail in [24].

Next, following [9], consider totally nonnegative matrices A. For such A, the exact LU factors have nonnegative elements, and the same is true of the computed ones if the unit roundoff is sufficiently small. In this case we have, using (3.9),

$$|\widehat{L}||\widehat{U}| = |\widehat{L}\widehat{U}| = |A + E| \leq |A| + \gamma_n|\widehat{L}||\widehat{U}|,$$

that is,

$$|\widehat{L}||\widehat{U}| \leq \frac{1}{1-\gamma_n}|A|.$$

Hence the backward error matrix E in (3.5) satisfies

$$|E| \leq \frac{\gamma_n(2+\gamma_n)}{1-\gamma_n}|A|,$$

that is, the componentwise relative backward error is pleasantly small. The same is true for some important classes of tridiagonal matrix. Higham [22] shows that if A is tridiagonal and either

- symmetric positive definite,
- an M-matrix ($a_{ij} \leq 0$ for all $i \neq j$ and $A^{-1} \geq 0$), or
- diagonally dominant by columns or by rows,

then

$$(A + E)\widehat{x} = b, \qquad |E| \leq f(u)|A|,$$

where $f(u) = cu + O(u^2)$, c being a constant of order unity.

Note that in the above examples row interchanges in the LU factorization destroy the required matrix properties and so we lose the favourable backward error bounds. In these cases it is advantageous *not* to pivot!

5 Computing Backward Error Bounds

In cases where it is not known a priori that the backward error is sufficiently small it is desirable to obtain an a posteriori bound for it. Several authors have considered how to compute or estimate the bounds of section 3.

Consider first Wilkinson's bound (3.2), in which the only "nontrivial" term is the growth factor ρ_n. The growth factor can be computed by monitoring the size of elements during the elimination, at a cost of $O(n^3)$ comparisons. This has been regarded as rather expensive, and more efficient ways to estimate ρ_n have been sought.

Businger [5] describes a way to obtain an upper bound for ρ_n in $O(n^2)$ operations. This approach is generalized by Erisman and Reid [14] who apply the Holder inequality to the equation

$$a_{ij}^{(k)} = a_{ij} - \sum_{r=1}^{k} l_{ir} u_{rj}, \qquad i, j > k,$$

to obtain the bound

$$\begin{aligned} |a_{ij}^{(k)}| &\leq |a_{ij}| + \|(l_{i1}, \dots, l_{ik})\|_p \|(u_{1j}, \dots, u_{kj})\|_q, \\ &\leq \max_{i,j} |a_{ij}| + \max_i \|(l_{i1}, \dots, l_{i,i-1})\|_p \max_j \|(u_{1j}, \dots, u_{j-1,j})\|_q, \end{aligned} \qquad (5.1)$$

where $p^{-1} + q^{-1} = 1$. In practice, $p = 1, 2, \infty$ are the values of interest. Barlow [2] notes that application of the Holder inequality instead to

$$a_{ij}^{(k)} = \sum_{r=k+1}^{\min(i,j)} l_{ir} u_{rj}$$

yields the sometimes sharper bound

$$|a_{ij}^{(k)}| \leq \max_i \|L(i,:)\|_p \max_j \|U(:,j)\|_q,$$

where $L(i,:)$ is the ith row of L and $U(:,j)$ is the jth column of U.

It is interesting to note that in the light of experience with the bound (5.1) Reid [33] recommends computing the growth factor explicitly in the context of sparse matrices, arguing that the expense is justified because (5.1) can be a very weak bound. See [13] for some empirical results on the quality of the bound.

Chartres and Geuder [6] propose computing the n^2 bounds in (3.3) explicitly. They note that since the terms $\widehat{l}_{ik}\widehat{u}_{kj}$ which make up W are formed anyway during the LU factorization, W can be computed in parallel with the factorization at a cost of $n^3/3$ additions.

Chu and George [7] observe that the ∞-norm of the matrix $|\widehat{L}||\widehat{U}|$ can be computed in $O(n^2)$ operations without forming the matrix explicitly, since

$$\| \, |\widehat{L}||\widehat{U}| \, \|_\infty = \| \, |\widehat{L}||\widehat{U}|e \, \|_\infty = \| \, |\widehat{L}|(|\widehat{U}|e) \, \|_\infty.$$

Thus one can cheaply compute a bound on $\|E\|_\infty$ from a componentwise backward error bound such as (3.5).

All the methods discussed above make use of an a priori error analysis to compute bounds on the backward error. Because the bounds do not take into account the statistical distribution of rounding errors, and because they have somewhat pessimistic constant terms, they cannot be expected to be very sharp. Thus it is important not to forget that, as shown is section 2, it is straightforward to compute the backward error itself! To obtain the backward error in the LU factorization we have to compute $\widehat{L}\widehat{U}$, which costs $O(n^3)$ operations. If the normwise backward error is wanted then one could instead *estimate* $\|A - \widehat{L}\widehat{U}\|_1$ in $O(n^2)$ operations using the matrix norm estimator described in section 7. The backward error for $\widehat{x}$ can be computed in $O(n^2)$ operations via just one or two matrix-vector products, from the formulas in section 2. We discuss the effect of rounding error on this computation in section 7.

6 Iterative Refinement

What can we do if the computed solution $\widehat{x}$ to $Ax = b$ does not have a small enough backward error? The traditional answer to a slightly different question—what to do if $\widehat{x}$ does not have a small enough forward error—is to use iterative refinement, and it is usually stressed that residuals must be computed in higher precision (see, e.g., [16]). Work by Jankowski and Woźniakowski [28] and Skeel [36] shows that iterative refinement using *single precision* residuals is usually sufficient to yield a small backward error, although it will not necessarily produce a small forward error. Jankowski and Woźniakowski deal with normwise backward error in their analysis and cater for arbitrary linear equation solvers. Skeel specializes to GEPP and uses the stronger componentwise relative backward error. Skeel's analysis and results are intricate, but the gist may be stated simply:

> If the product of $\mathrm{cond}(A^{-1}) \equiv \|\,|A||A^{-1}|\,\|_\infty$ and $\sigma(A,x) \equiv \max_i(|A||x|)_i/\min_i(|A||x|)_i$ is less than $(f(A,b)u)^{-1}$, where $f(A,b)$ is typically $O(n)$, then after GEPP with one step of single precision iterative refinement $\beta_C(|A|,|b|) \le (n+1)u$.

Thus after just one step of iterative refinement in single precision GEPP—which is already stable in the sense of having small normwise backward error—is much more strongly stable: it has a small componentwise relative backward error, provided that the problem is not too ill-conditioned ($\mathrm{cond}(A^{-1})$ is not too large) or too badly scaled ($\sigma(A,x)$ is not too large).

Arioli, Demmel and Duff [1] consider in detail the practical use of iterative refinement in single precision with GEPP, with an emphasis on sparse matrices. They note that $\sigma(A,x)$ can be very large for systems in which both A and b are sparse, in which case Skeel's result is not applicable. To circumvent this problem they change the backward error measure from $\beta_C(|A|,|b|)$ to $\beta_C(|A|,f)$, where f is chosen in an a posteriori way in which f_i is permitted to exceed $|b_i|$. Thus in the backward error definition the sparsity of A is preserved while that of b may be sacrificed in order to make the backward error small after one or more steps of iterative refinement. See [1] for further details, and a comprehensive suite of numerical tests. For a thorough survey of iterative refinement in linear equations and other contexts see [3].

7 Estimating and Bounding the Forward Error

The usual way to explore the forward error $\|x-\hat{x}\|$ is by applying perturbation theory to a backward error analysis. Corresponding to the normwise backward error result

$$(A+\Delta A)\hat{x} = b+\Delta b, \qquad \|\Delta A\| \le \epsilon\|A\|, \quad \|\Delta b\| \le \epsilon\|b\|,$$

we have the well-known bound

$$\frac{\|x-\hat{x}\|}{\|x\|} \le \frac{2\epsilon\kappa(A)}{1-\epsilon\kappa(A)} \qquad (\epsilon\kappa(A)<1), \tag{7.1}$$

where the condition number $\kappa(A) = \|A\|\|A^{-1}\|$. For the componentwise analysis

$$(A+\Delta A)\hat{x} = b+\Delta b, \qquad |\Delta A| \le \omega E, \quad |\Delta b| \le \omega f,$$

a straightforward generalization of a result in [35, Theorem 2.1] yields

$$\frac{\|x-\hat{x}\|_\infty}{\|x\|_\infty} \le \frac{\omega\kappa_{E,f}(A,b)}{1-\omega\kappa_E(A)} \qquad (\omega\kappa_E(A)<1), \tag{7.2}$$

where, in the notation of [1],

$$\kappa_{E,f}(A,b) \equiv \frac{\|\,|A^{-1}|E|x| + |A^{-1}|f\,\|_\infty}{\|x\|_\infty},$$
$$\kappa_E(A) \equiv \|\,|A^{-1}|E\,\|_\infty.$$

Of particular interest is the condition number for the componentwise relative backward error, $\kappa_{|A|,|b|}(A,b)$. This is easily seen to differ by no more than a factor 2 from, using Skeel's notation [35],

$$\operatorname{cond}(A,x) \equiv \frac{\|\,|A^{-1}||A||x|\,\|_\infty}{\|x\|_\infty}.$$

The maximum value of $\mathrm{cond}(A,x)$ occurs for $x = e$ and is

$$\mathrm{cond}(A) \equiv \| \, |A^{-1}||A| \, \|_\infty.$$

It is an important fact that $\mathrm{cond}(A)$ is never bigger than $\kappa_\infty(A)$, and it can be much smaller. This is because $\mathrm{cond}(A,x)$ is independent of the row scaling of A while $\kappa_\infty(A)$ is not. One implication of this row scaling independence is that $\mathrm{cond}(A)$ may differ greatly from $\mathrm{cond}(A^T)$. Thus in the sense of componentwise perturbations $Ax = b$ can be much more or less ill-conditioned than $A^T y = c$ (see [24] for some specific examples). Another implication is that (7.2) can sometimes provide a much smaller upper bound than (7.1). Note also that the forward error depends on the right-hand side b, and that (7.2) displays this dependency but (7.1) does not.

We note that under assumptions on the statistical distribution of the perturbations ΔA and Δb an expression can be derived for the expected 2-norm of the forward error [15]; this may be preferable to (7.1) and (7.2) when statistical information about ΔA and Δb is available.

Both bounds (7.1) and (7.2) contain a condition number involving A^{-1}. In most cases exact computation of the condition number would necessitate forming A^{-1}. To avoid this expense it is standard practice to compute an inexpensive estimate of the condition number using a condition estimator. Various condition estimation techniques have been developed—see [21] for a survey. The most well-known estimator is the one used in LINPACK, which provides a lower bound for $\kappa_1(A)$. The method underlying this estimator does not generalize to the estimation of $\kappa_{E,f}(A,b)$. A more versatile estimator with this capability is one developed by Hager [20] and Higham [23]. This estimator treats the general problem of estimating $\|B\|_1$, where B is not known explicitly. The estimator assumes that B is described by a "black box" that can evaluate Bx or $B^T x$ given x. Typically, 4 or 5 such matrix-vector products are required to produce a lower bound for $\|B\|_1$, and the bound is almost invariably within a factor 3 of $\|B\|_1$. To estimate $\kappa_1(A)$ we take $B = A^{-1}$, and the estimator requires the solution of linear systems with A and A^T as coefficient matrices, which is inexpensive if an LU factorization of A is available. Arioli, Demmel and Duff [1] show how to apply the estimator to $\kappa_{E,f}(A,b)$. The problem is basically that of estimating $\||A^{-1}|g\|_\infty$, where $g \geq 0$. With $G = \mathrm{diag}(g_1, g_2, \ldots, g_n)$ the equalities

$$\| \, |A^{-1}|g \, \|_\infty = \| \, |A^{-1}|Ge \, \|_\infty = \| \, |A^{-1}G|e \, \|_\infty = \| \, |A^{-1}G| \, \|_\infty = \|A^{-1}G\|_\infty$$

show that the problem reduces to estimating $\|B\|_1$ where $B = (A^{-1}G)^T$ and where Bx and $B^T y$ can be formed by solving linear systems involving A^T and A respectively.

There are certain circumstances in which one can compute $\kappa_{E,f}(A,b)$ *exactly* with the same order of work as solving a linear system given an LU factorization of A. Higham [22] shows that if the nonsingular matrix A is tridiagonal and nonsingular and has an LU factorization with $|L||U| = |A|$ then

$$|U^{-1}||L^{-1}| = |A^{-1}|. \tag{7.3}$$

Recall from section 4.2 that the condition $|L||U| = |A|$ implies a small componentwise relative backward error for GE; it holds when the tridiagonal matrix is symmetric positive definite, totally nonnegative, or an M-matrix [22]. To see the significance of (7.3) consider a 3×3 bidiagonal matrix and its inverse:

$$U = \begin{bmatrix} d_1 & e_2 & \\ & d_2 & e_3 \\ & & d_3 \end{bmatrix}, \qquad U^{-1} = \begin{bmatrix} d_1^{-1} & -e_2 d_1^{-1} d_2^{-1} & e_2 e_3 d_1^{-1} d_2^{-1} d_3^{-1} \\ & d_2^{-1} & -e_3 d_2^{-1} d_3^{-1} \\ & & d_3^{-1} \end{bmatrix}.$$

Clearly we have

$$|U^{-1}| = M(U)^{-1} \qquad \text{where} \quad M(U) = \begin{bmatrix} |d_1| & -|e_2| & \\ & |d_2| & -|e_3| \\ & & |d_3| \end{bmatrix}$$

($M(U)$ is called the comparison matrix for U). The relation $|B^{-1}| = M(B)^{-1}$ is true for all bidiagonal B, so

$$|A^{-1}|g = |U^{-1}||L^{-1}|g = M(U)^{-1}M(L)^{-1}g,$$

which may be computed by solving two bidiagonal systems. Thus for the three classes of tridiagonal matrix mentioned above it costs only $O(n)$ operations to compute $\kappa_{E,f}(A, b)$ given an LU factorization of A. The same technique can be used for M-matrices in general, because $A^{-1} \geq 0$ implies $|A^{-1}|g = A^{-1}g$.

Next we describe further details of estimating the forward error in practice. To evaluate bounds (7.1) and (7.2) we need the backward error, or a bound for it. We could use the backward error bounds from section 3, but as explained in section 5 it is better to compute the backward error directly using the formulas from section 2.

A more direct approach is to use the following bound, which is so straightforward that it is easily overlooked: from $x - \hat{x} = A^{-1}r$, where $r = b - A\hat{x}$,

$$\|x - \hat{x}\|_\infty \leq \|\, |A^{-1}||r|\, \|_\infty.$$

Note that the inequality arises solely from ignoring signs of terms in the matrix-vector product. Assuming r is computed in single precision, the rounding errors in its formation are accounted for by (using a variation of Lemma 3.2)

$$\hat{r} = r + \Delta r, \qquad |\Delta r| \leq \gamma_{n+1}(|b| + |A||\hat{x}|).$$

Thus, our final practical bound, which we estimate using the norm estimator described above, is

$$\frac{\|x - \hat{x}\|_\infty}{\|\hat{x}\|_\infty} \leq \frac{\|\, |A^{-1}|(|\hat{r}| + \gamma_{n+1}(|b| + |A||\hat{x}|))\, \|_\infty}{\|\hat{x}\|_\infty}. \tag{7.4}$$

Note that the $|b| + |A||\hat{x}|$ term needs to be computed anyway if we are evaluating the componentwise relative backward error from (2.4).

Arioli, Demmel and Duff [1] found in their experiments that after iterative refinement (7.4) gave similar sized bounds to (7.2) with $E = |A|$, $f = |b|$, and with ω computed a posteriori using (2.4).

We conclude this section by noting the need for care in interpreting the forward error. Experiments in [24] show that simply changing the order of evaluation of an inner product in the substitution algorithm for solution of a triangular system can change the forward error in the computed solution by orders of magnitude. This means, for example, that it is dangerous to compare different codes or algorithms solely in terms of observed forward errors.

8 Concluding Remarks

The work we have described falls broadly into two areas:

- normwise error analysis and iterative refinement in double precision, and
- componentwise analysis and iterative refinement in single precision.

The first area might be described as being "traditional" or "in the style of Wilkinson", while the second has been the subject of most of the recent research in this subject.

An indication of the success of the recent work is that LAPACK [11], the successor to LINPACK and EISPACK that is currently under development, will incorporate iterative refinement in single precision with computation of the componentwise relative backward error and estimation of $\kappa_{|A|,|b|}(A, b)$ (see [12]).

Finally, we return to the numerical example in section 1. The Vandermonde matrix A in this example is totally nonnegative, so we know from section 4.2 that the componentwise relative backward error must be small for GE. For GEPP there is no such guarantee since the permuted matrix "PA" is not totally nonnegative—as we saw, only the normwise backward error is small in this example. We have $\kappa_\infty(A) \approx 4 \times 10^7$, and $\mathrm{cond}(A, x) \approx 2 \times 10^4$ ($\mathrm{cond}(A) \approx 9 \times 10^4$). The forward errors for the two computed solutions are consistent with the "forward error $\leq$ condition number $\times$ backward error" results (7.1) and (7.2). Concerning iterative refinement in single precision, the product $\mathrm{cond}(A^{-1})\sigma(A, x) \approx 10^3 u$, so Skeel's result quoted in section 6 is not applicable. Nevertheless, the first iterate has a small componentwise relative backward error and its forward error is consistent with the forward error bounds.

Acknowledgements

I thank Des Higham and Steve Vavasis for carefully reading the manuscript and suggesting numerous improvements.

REFERENCES

[1] M. Arioli, J.W. Demmel and I.S. Duff, Solving sparse linear systems with sparse backward error, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 165–190.

[2] J.L. Barlow, A note on monitoring the stability of triangular decomposition of sparse matrices, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 166–168.

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

[4] J.R. Bunch, J.W. Demmel and C.F. Van Loan, The strong stability of algorithms for solving symmetric linear systems; to appear in SIAM J. Matrix Anal. Appl.

[5] P.A. Businger, Monitoring the numerical stability of Gaussian elimination, Numer. Math., 16 (1971), pp. 360–361.

[6] B.A. Chartres and J.C. Geuder, Computable error bounds for direct solution of linear equations, J. Assoc. Comput. Mach., 14 (1967), pp. 63–71.

[7] E. Chu and A. George, A note on estimating the error in Gaussian elimination without pivoting, ACM SIGNUM Newsletter, 20 (1985), pp. 2–7.

[8] S.D. Conte and C. de Boor, *Elementary Numerical Analysis*, Third Edition, McGraw-Hill, Tokyo, 1980.

[9] C. de Boor and A. Pinkus, Backward error analysis for totally positive linear systems, Numer. Math., 27 (1977), pp. 485–490.

[10] J.W. Demmel, Underflow and the reliability of numerical software, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 887–919.

[11] J.W. Demmel, J.J. Dongarra, J.J. Du Croz, A. Greenbaum, S.J. Hammarling and D.C. Sorensen, Prospectus for the development of a linear algebra library for high-performance computers, Technical Memorandum No. 97, Mathematics and Computer Science Division, Argonne National Laboratory, Illinois, 1987.

[12] J.W. Demmel, J.J. Du Croz, S.J. Hammarling and D.C. Sorensen, Guidelines for the design of symmetric eigenroutines, SVD, and iterative refinement and condition estimation for linear systems, LAPACK Working Note #4, Technical Memorandum 111, Mathematics and Computer Science Division, Argonne National Laboratory, 1988.

[13] 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. Stat. Comput., 8 (1987), pp. 600–624.

[14] 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.

[15] R. Fletcher, Expected conditioning, IMA Journal of Numerical Analysis, 5 (1985), pp. 247–273.

[16] G.E. Forsythe and C.B. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

[17] L. Fox, *An Introduction to Numerical Linear Algebra*, Oxford University Press, 1964.

[18] L. Fox, H.D. Huskey and J.H. Wilkinson, Notes on the solution of algebraic linear simultaneous equations, Quart. J. Mech. and Applied Math., 1 (1948), pp. 149–173.

[19] G.H. Golub and C.F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, Maryland, 1983.

[20] W.W. Hager, Condition estimates, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 311–316.

[21] N.J. Higham, A survey of condition number estimation for triangular matrices, SIAM Review, 29 (1987), pp. 575–596.

[22] N.J. Higham, Bounding the error in Gaussian elimination for tridiagonal systems, Technical Report 88-953, Department of Computer Science, Cornell University, 1988.

[23] N.J. Higham, FORTRAN codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation, ACM Trans. Math. Soft., 14 (1988), pp. 381–396.

[24] N.J. Higham, The accuracy of solutions to triangular systems, Numerical Analysis Report No. 158, University of Manchester, England, 1988; to appear in SIAM J. Numer. Anal.

[25] N.J. Higham, Analysis of the Cholesky decomposition of a semi-definite matrix, in *Reliable Numerical Computation*, M.G. Cox and S.J. Hammarling, eds., Oxford University Press, 1989.

[26] N.J. Higham and D.J. Higham, Large growth factors in Gaussian elimination with pivoting, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 155–164.

[27] H. Hotelling, Some new methods in matrix calculation, Ann. Math. Statist., 14 (1943), pp. 1–34.

[28] M. Jankowski and H. Woźniakowski, Iterative refinement implies numerical stability, BIT, 17 (1977), pp. 303–311.

[29] W. Oettli and W. Prager, Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides, Numer. Math., 6 (1964), pp. 405–409.

[30] F.W.J. Olver and J.H. Wilkinson, *A posteriori* error bounds for Gaussian elimination, IMA Journal of Numerical Analysis, 2 (1982), pp. 377–406.

[31] G. Peters and J.H. Wilkinson, On the stability of Gauss-Jordan elimination with pivoting, Comm. ACM, 18 (1975), pp. 20–24.

[32] J.K. Reid, A note on the stability of Gaussian elimination, J. Inst. Maths. Applics., 8 (1971), pp. 374–375.

[33] J.K. Reid, Sparse matrices, in *The State of the Art in Numerical Analysis*, A. Iserles and M.J.D. Powell, eds., Oxford University Press, 1987, pp.

59–85.
[34] J.L. Rigal and J. Gaches, On the compatibility of a given solution with the data of a linear system, J. Assoc. Comput. Mach., 14 (1967), pp. 543–548.
[35] R.D. Skeel, Scaling for numerical stability in Gaussian elimination, J. Assoc. Comput. Mach., 26 (1979), pp. 494–526.
[36] R.D. Skeel, Iterative refinement implies numerical stability for Gaussian elimination, Math. Comp., 35 (1980), pp. 817–832.
[37] R.D. Skeel, Effect of equilibration on residual size for partial pivoting, SIAM J. Numer. Anal., 18 (1981), pp. 449–454.
[38] G.W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
[39] G.W. Stewart, Research, development, and LINPACK, in *Mathematical Software III*, J.R. Rice, ed., Academic Press, New York, 1977, pp. 1–14.
[40] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.
[41] F. Stummel, Forward error analysis of Gaussian elimination, Part I: Error and residual estimates, Numer. Math., 46 (1985), pp. 365–395.
[42] F. Stummel, Forward error analysis of Gaussian elimination, Part II: Stability theorems, Numer. Math., 46 (1985), pp. 397–415.
[43] L.N. Trefethen and R.S. Schreiber, Average-case stability of Gaussian elimination, Numerical Analysis Report 88-3, Department of Mathematics, M.I.T., 1988; to appear in SIAM J. Matrix Anal. Appl.
[44] A.M. Turing, Rounding-off errors in matrix processes, Quart. J. Mech. and Applied Math., 1 (1948), pp. 287–308.
[45] A. van der Sluis, Condition, equilibration and pivoting in linear algebraic systems, Numer. Math., 15 (1970), pp. 74–86.
[46] J. von Neumann and H.H. Goldstine, Numerical inverting of matrices of high order, Bull. Amer. Math. Soc., 53 (1947), pp. 1021–1099.
[47] B. Wendroff, *Theoretical Numerical Analysis*, Academic Press, New York, 1966.
[48] J.H. Wilkinson, Error analysis of direct methods of matrix inversion, J. Assoc. Comput. Mach., 8 (1961), pp. 281–330.
[49] J.H. Wilkinson, *Rounding Errors in Algebraic Processes*, Notes on Applied Science No. 32, Her Majesty's Stationery Office, London, 1963.
[50] J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, 1965.
[51] J.H. Wilkinson, Modern error analysis, SIAM Review, 13 (1971), pp. 548–568.
[52] J.H. Wilkinson, The state of the art in error analysis, NAG Newsletter 2/85, Numerical Algorithms Group, Oxford, 1985.
[53] J.H. Wilkinson, Error analysis revisited, IMA Bulletin, 22 (1987), pp. 192–200.