SIAM J. NUMER. ANAL.
Vol. 26, No. 5, pp. 1252-1265, October 1989

013

THE ACCURACY OF SOLUTIONS TO TRIANGULAR SYSTEMS*

NICHOLAS J. HIGHAM†

Abstract. Triangular systems play a fundamental role in matrix computations. It has been prominently stated in the literature, but is perhaps not widely appreciated, that solutions to triangular systems are usually computed to high accuracy—higher than the traditional condition numbers for linear systems suggest. This phenomenon is investigated by use of condition numbers appropriate to the componentwise backward error analysis of triangular systems. Results of Wilkinson are unified and extended. Among the conclusions are that the conditioning of a triangular system depends on the right-hand side as well as the coefficient matrix; that use of pivoting in LU, QR, and Cholesky factorisations can greatly improve the conditioning of a resulting triangular system; and that a triangular matrix may be much more or less ill-conditioned than its transpose.

Key words. triangular matrix, triangular system, substitution algorithm, forward error analysis, backward error analysis, condition number, comparison matrix, M-matrix, pivoting

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

CR subject classification. G.1.3

1. Introduction. Triangular matrices are ubiquitous in matrix computations. Their importance is due to the fact that practically all direct methods (and many iterative methods) for solving general linear systems involve the solution of triangular systems, which is easily done using the standard back and forward substitution algorithms. Since triangular systems play such a fundamental role in matrix computations, it is desirable to understand fully their solution in floating-point arithmetic. Although other methods have been devised for solving triangular systems [11], the substitution algorithms are universally used, and we concentrate on these algorithms here.

The backward error analysis for solution of a triangular system is straightforward and well known. In contrast, the behaviour of the forward error is rarely discussed. We might assume that, as is true for general linear systems, we can obtain useful forward error bounds and error estimates from the backward error analysis by applying standard perturbation theory involving the matrix condition number. However, statements in three classic texts in matrix computations call this assumption into question. Wilkinson states in [21, p. 105] and makes a similar statement in [22, p. 251], "In practice one almost invariably finds that if L is ill-conditioned, so that $\|L\|\,\|L^{-1}\| \gg 1$, then the computed solution of $Lx = b$ (or the computed inverse) is far more accurate than [standard norm bounds] would suggest." Likewise, Stewart [18, p. 150] explains:

> *The solutions of triangular systems are usually computed to high accuracy.* This fact · · · cannot be proved in general, for counter examples exist. However, it is true of many special kinds of triangular matrices and the phenomenon has been observed in many others. The practical consequences of this fact cannot be over-emphasized.

These clear and prominent statements are supported, although not completely explained, by analysis given by Wilkinson [20], [22]; we summarise this analysis in § 2. Surprisingly, no further analysis seems to have been published, although empirical observations of high-accuracy solutions of triangular systems are reported in [8] and [12].

* Received by the editors July 25, 1988; accepted for publication (in revised form) December 3, 1988.
† Department of Mathematics, University of Manchester, Manchester M13 9PL, United Kingdom. Present address, Department of Computer Science, Cornell University, Ithaca, New York 14853.

The purpose of this paper is twofold. First, in § 3 we present a unified derivation of Wilkinson's results, extending some and phrasing all in terms of floating-point arithmetic. The key tools are the componentwise perturbation theory and associated condition numbers of Skeel [17]. Second, in § 4 we present the results of numerical experiments designed to confirm and illustrate the analysis, and to give further insight into the numerical behaviour in practice.

Of course, in most applications, solving a triangular system forms just part of an algorithm, and even if the triangular system is solved *exactly*, we may not be able to draw stronger conclusions about the error properties of the overall algorithm. Nevertheless, it is interesting and useful to know when and why triangular systems are solved to "high accuracy," and in precisely what sense. We can draw an analogy with [3], where, concerning the second stage of the SVD algorithm, high-accuracy computation of singular values of bidiagonal matrices is considered.

We stress that the analysis given here is applicable to all forms of the substitution algorithms: the inner product ("SDOT") and vector sum ("SAXPY") orderings [4], implemented either in the standard serial fashion or in the sophisticated parallel versions that have recently been developed (see [10] for a survey). Fortran software for the substitution algorithms is widely available, notably in LINPACK [4] and in the Level 2 BLAS [5], [6].

2. Wilkinson's analysis. In this section we summarise Wilkinson's results on error analysis for triangular systems.

Recall that for an $n \times n$ upper triangular system $Ux = b$ the back substitution algorithm is

$$\text{for } i = n, n-1, \cdots, 1$$

$$s = \sum_{j=i+1}^{n} u_{ij} x_j$$

$$x_i = \frac{b_i - s}{u_{ii}}$$

$$\text{endfor}$$

where the empty sum is defined to be zero. The following backward error result assumes that the algorithm is carried out in precisely the manner indicated (and similarly for forward substitution). We stress that the backward error bound depends on the order in which the terms in the inner product are accumulated, and on the stage at which b_i is added to the sum. However, the simple expedient of replacing the ordering-dependent term $|i-j|+2$ in (2.2) below by $n+1$ (as we shall do in § 3) makes the result applicable to all implementations of the substitution process.

Assume that computations are carried out in a floating-point arithmetic that obeys the model

$$fl(x \text{ op } y) = (x \text{ op } y)(1+\delta), \quad |\delta| \leqq u, \quad \text{op} = +, -, *, /,$$

where u is the unit roundoff. (In fact, the following result holds under a weaker model encompassing machines that do not use a guard digit in addition or subtraction.) Assume also that floating-point underflow or overflow does not occur.

THEOREM 2.1 [21, p. 100], [18, pp. 150, 408]. *Let $T \in \mathbf{R}^{n \times n}$ be a nonsingular triangular matrix, and assume $nu < 0.1$. Then the computed solution $\hat{x}$ to the system $Tx = b$ satisfies*

$$(T+E)\hat{x} = b, \tag{2.1}$$

where

$$|e_{ij}| \leq (|i-j|+2)cu|t_{ij}|, \qquad 1 \leq i, j \leq n, \tag{2.2}$$

in which c is a constant of order unity.

The theorem shows that $\hat{x}$ is the exact solution of a system obtained from $Tx = b$ by making small componentwise relative perturbations to T. Note that the zero elements of T are not perturbed.

The usual perturbation analysis proceeds by obtaining from (2.1) the bound

$$\frac{\|x-\hat{x}\|}{\|x\|} \leq \frac{\kappa(T)\|E\|/\|T\|}{1-\kappa(T)\|E\|/\|T\|} \qquad (\kappa(T)\|E\|/\|T\| < 1),$$

where the condition number $\kappa(T) = \|T\|\,\|T^{-1}\|$. For the $1, \infty$, and Frobenius norms, (2.2) implies

$$\|E\| \leq c_n u \|T\|, \qquad c_n = (n+1)c,$$

and so we have for these norms

$$\frac{\|x-\hat{x}\|}{\|x\|} \leq \frac{\kappa(T)c_n u}{1-\kappa(T)c_n u} \qquad (\kappa(T)c_n u < 1). \tag{2.3}$$

In [20]-[22] Wilkinson notes that the bound (2.3) is often very pessimistic, and he shows that in certain cases much stronger forward error bounds can be derived. We now summarise Wilkinson's forward error results.

Let L denote an $n \times n$ lower triangular matrix. Results (1)-(3) are concerned with computation of $X = L^{-1}$. $\hat{X}$ denotes the computed inverse, whose ith column is obtained by solving the system $Lx_i = e_i$, where e_i is the ith column of the identity matrix. Results (4) and (5) are concerned with solution of a single triangular system $Ly = b$.

(1) [20, p. 322]. If L has positive diagonal elements and nonpositive off-diagonal elements, then

$$|x_{ij} - \hat{x}_{ij}| \leq 3(i-j+1)u|x_{ij}|, \qquad 1 \leq j \leq i \leq n.$$

Thus every element of the computed inverse has a small relative error, independent of the condition of L.

The next two results assume the use of fixed-point arithmetic with a constant scale factor.

(2) [20, p. 323]. If $|l_{ii}| \geq |l_{ij}|$ for all $j < i$, then

$$|x_{ij} - \hat{x}_{ij}| \leq 2^{i-j+1} u \max_k |x_{kj}|, \qquad 1 \leq j \leq i \leq n.$$

Wilkinson comments, "Hence if we have used complete pivoting on a matrix of lower order we are certain to get a 'comparatively good' result when the matrix is ill-conditioned."

(3) [20, p. 324-325]. If X satisfies $|x_{ij}| \leq \theta|x_{jj}|$ for $i > j$, then

$$|x_{ij} - \hat{x}_{ij}| \leq \theta n u \max_{r,s} |\hat{x}_{rs}|, \qquad 1 \leq j \leq i \leq n.$$

The final two results assume the use of floating-point arithmetic with double-length accumulation of inner products. Wilkinson states [22, p. 250] "For other forms of

computation the upper bounds obtained are somewhat poorer but the same broad features persist."

(4) [22, p. 249]. If X satisfies $|x_{ij}| \leqq \theta |x_{jj}|$ for $i > j$, then

$$\frac{\|y - \hat{y}\|_\infty}{\|y\|_\infty} \leqq \frac{\theta n u}{1 - \theta n u}.$$

Wilkinson notes that θ is often of order unity when L is very ill-conditioned.

(5) [22, p. 250]. If L has positive diagonal elements and nonpositive off-diagonal elements, and b has nonnegative elements, then

$$|y_i - \hat{y}_i| \leqq \varepsilon_i |y_i|,$$

where

$$\varepsilon_i = (1+u)^i (1+\tfrac{3}{2}u^2)^{i(i+1)/2} - 1.$$

3. The forward error. Our goal in this section is to present a unified derivation of Wilkinson's results, phrasing them all in terms of floating-point arithmetic, and extending the results where possible.

Let $|\cdot|$ denote the operation of replacing each element of a vector or matrix by its absolute value. For the following analysis we need to write the backward error result of Theorem 2.1 in the form

$$(T+E)\hat{x} = b, \tag{3.1a}$$

$$|E| \leqq c_n u |T|, \tag{3.1b}$$

where $c_n = (n+1)c$. This represents a weakening of (2.2) for most elements of E, but, as noted earlier, it makes the analysis valid for all implementations of the substitution process.

Note that by partitioning (3.1) in the form (for upper triangular T)

$$\begin{pmatrix} T_{11}+E_{11} & T_{12}+E_{12} \\ 0 & T_{22}+E_{22} \end{pmatrix} \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

we obtain

$$(T_{22}+E_{22})\hat{x}_2 = b_2, \qquad |E_{22}| \leqq c_n u |T_{22}|,$$

and the analysis below can be applied to any such subsystem to obtain bounds on the error in $\hat{x}_2$ that are independent of $\hat{x}_1$. This can be useful if $\|\hat{x}_1\|_\infty \gg \|\hat{x}_2\|_\infty$.

To analyse (3.1) we make use of relevant perturbation theory and condition numbers of Skeel [17]. Skeel considers general square linear systems $Ax = b$ subject to perturbations $A \to A + E$, $|E| \leqq \varepsilon |A|$, and $b \to b + d$, $|d| \leqq \varepsilon |b|$. For perturbations in A alone Skeel introduces the condition number

$$\operatorname{cond}(A, x) \equiv \lim_{\varepsilon \to 0} \sup_{|E| \leqq \varepsilon |A|} \frac{\|\delta x\|_\infty}{\varepsilon \|x\|_\infty},$$

where $(A+E)(x+\delta x) = b$, and shows that

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

The maximum value of $\operatorname{cond}(A, x)$ is

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

It is straightforward to derive from (3.1) the bound (see, for example, [17, Thm. 2.1])

$$\frac{\|x - \hat{x}\|_\infty}{\|x\|_\infty} \leqq \frac{\operatorname{cond}(T, x) c_n u}{1 - \operatorname{cond}(T) c_n u} \qquad (\operatorname{cond}(T) c_n u < 1). \tag{3.2}$$

Skeel's theory tells us that for any T, $\hat{x}$, and b, there exists an E satisfying (3.1) for which there is approximate equality in (3.2). Thus (3.2) must be regarded as a sharp bound in practice (unlike (2.3)), and so it is appropriate to concentrate our efforts on assessing the size of cond (T, x).

The most important feature of cond (T, x) is that it is invariant under row scaling of T; this follows from the relation, for $D = \operatorname{diag}(d_i)$,

$$|(DT)^{-1}||DT| = |T^{-1}||D^{-1}||D||T| = |T^{-1}||T|. \tag{3.3}$$

The underlying reason for this invariance is that in (3.1) a row scaling of T and b is reflected in the bound for $|E|$, and thus (3.1) is essentially left unchanged by such a scaling.

In terms of the traditional condition number $\kappa(T)$, ill-conditioning of a triangular matrix stems from two possible sources: variation in the size of the diagonal elements, and rows with off-diagonal elements that are large relative to the diagonal element. Significantly, because of the row scaling invariance, cond (T, x) is susceptible only to the second source. An extreme example of the difference between cond (T) and $\kappa(T)$ is the case of diagonal matrices D: $\kappa(D)$ can be arbitrarily large, yet cond $(D) \equiv 1$.

Despite its pleasing properties, cond (T, x) can be arbitrarily large. This is illustrated by the upper triangular matrix

$$T(\alpha) = (t_{ij}), \qquad t_{ij} = \begin{cases} 1, & i = j, \\ -\alpha, & i < j, \end{cases} \tag{3.4}$$

for which cond $(T(\alpha), e) = \text{cond}\,(T(\alpha)) \sim 2\alpha^{n-1}$ as $\alpha \to \infty$. Therefore we cannot assert that *all* triangular systems are solved to high accuracy. Nevertheless, for any T there is always at least one system for which high accuracy is obtained: the system $Tx = e_1$ if T is upper triangular, or $Tx = e_n$ if T is lower triangular. In both cases cond $(T, x) = 1$, and the solution comprises the computation of just a single scalar reciprocal.

To gain further insight we consider special classes of triangular matrices. In all the results below, T is assumed to be $n \times n$ and nonsingular.

LEMMA 3.1. *Suppose the upper triangular matrix T satisfies*

$$|t_{ii}| \geqq |t_{ij}| \quad \textit{for all } j > i. \tag{3.5}$$

Then the unit upper triangular matrix $W = |T^{-1}||T|$ satisfies $w_{ij} \leqq 2^{j-i}$ for all $j > i$.

Proof. By (3.3), $W = |U^{-1}||U|$, where $U = D^{-1}T$, $D = \operatorname{diag}(t_{ii})$. U is unit upper triangular with $|u_{ij}| \leqq 1$ for $j > i$, and it is easy to show that $|(U^{-1})_{ij}| \leqq 2^{j-i-1}$. Thus, for $j > i$,

$$w_{ij} = \sum_{k=i}^{j} |(U^{-1})_{ik}||u_{kj}| \leqq 1 + \sum_{k=i+1}^{j} 2^{k-i-1} \cdot 1 = 2^{j-i}. \qquad \square$$

THEOREM 3.2. *Under the conditions of Lemma 3.1 the computed solution $\hat{x}$ to $Tx = b$ satisfies*

$$|x_i - \hat{x}_i| \leqq 2^{n-i+1} c_n u \max_{j \geqq i} |\hat{x}_j|, \qquad 1 \leqq i \leqq n.$$

Proof. From (3.1) we have

$$|x - \hat{x}| = |T^{-1}E\hat{x}| \leqq c_n u |T^{-1}||T||\hat{x}|.$$

Using Lemma 3.1, we obtain

$$|x_i - \hat{x}_i| \leqq c_n u \sum_{j=i}^{n} w_{ij}|\hat{x}_j| \leqq c_n u \max_{j \geqq i} |\hat{x}_j| \sum_{j=i}^{n} 2^{j-i} \leqq 2^{n-i+1} c_n u \max_{j \geqq i} |\hat{x}_j|. \qquad \square$$

Lemma 3.1 shows that for matrices satisfying (3.5), cond (T) is bounded for fixed n no matter how large $\kappa(T)$. The bounds for $|x_i - \hat{x}_i|$ in Theorem 3.2, although large if n is large and i is small, decay exponentially with increasing i—thus later components of x are always computed to high accuracy relative to the elements already computed.

Analogues of Lemma 3.1 and Theorem 3.2 hold for lower triangular T satisfying

$$|t_{ii}| \geqq |t_{ij}| \quad \text{for all } j < i. \tag{3.6}$$

Note, however, that if the upper triangular matrix T satisfies (3.5), then T^T does not necessarily satisfy (3.6). In fact cond (T^T) can be arbitrarily large, as shown by the example

$$T = \begin{bmatrix} 1 & 1 & 0 \\ 0 & \varepsilon & \varepsilon \\ 0 & 0 & 1 \end{bmatrix},$$

$$\text{cond}\,(T) = 5, \qquad \text{cond}\,(T^T) = 1 + \frac{2}{\varepsilon}.$$

An important conclusion is that a triangular system $Tx = b$ can be much more or less ill-conditioned than the system $T^T y = c$, even if T satisfies (3.5).

Compared with Wilkinson's result (2) in § 2, Theorem 3.2 assumes floating-point arithmetic and is valid for any b rather than just the unit vectors. Theorem 3.2, or its lower triangular analogue, is applicable to:

- The lower triangular matrices from Gaussian elimination with partial pivoting or complete pivoting;
- The upper triangular matrices from Gaussian elimination with complete pivoting;
- The upper triangular matrices from the Cholesky and QR decompositions with complete pivoting and column pivoting, respectively.

Next, we consider triangular M-matrices, that is, triangular T satisfying

$$t_{ii} > 0, \quad t_{ij} \leqq 0 \quad \text{for all } i \neq j.$$

With a general triangular T there is associated an M-matrix called the comparison matrix:

$$M(T) = (m_{ij}), \qquad m_{ij} = \begin{cases} |t_{ii}|, & i = j, \\ -|t_{ij}|, & i \neq j. \end{cases}$$

The following result shows that, among all matrices R such that $|R| = |T|$, $R = M(T)$ is the one that maximises cond (R, x).

LEMMA 3.3. *For any triangular T,*

$$\text{cond}\,(T, x) \leqq \text{cond}\,(M(T), x) = \|(2M(T)^{-1}\,\text{diag}\,(|t_{ii}|) - I)|x|\,\|_\infty / \|x\|_\infty.$$

Proof. The inequality follows from $|T^{-1}| \leqq M(T)^{-1}$ [13], together with $|T| = |M(T)|$. We have

$$\begin{aligned} |M(T)^{-1}|\,|M(T)| &= M(T)^{-1}(2\,\text{diag}\,(|t_{ii}|) - M(T)) \\ &= 2M(T)^{-1}\,\text{diag}\,(|t_{ii}|) - I, \end{aligned}$$

which yields the equality. □

From the expression for cond $(M(T), x)$ in Lemma 3.3 it is easy to see that

$$\text{cond}\,(M(T), x) \leqq 1 + 2(n-1)\theta(M(T)),$$

where

$$\theta(T) \equiv \max_{i,j} \frac{|T^{-1}|_{ij}}{|T^{-1}|_{jj}} \leqq \operatorname{cond}(T).$$

Note that $\theta(T)$ is the quantity appearing in Wilkinson's bounds (3) and (4) of § 2. Unfortunately, it does not seem possible to obtain a useful bound for $\operatorname{cond}(T, x)$ in terms of $\theta(T)$.

An interesting feature of triangular M-matrices is that they exhibit two extremes of behaviour in the quantity $\operatorname{cond}(T, x)$. On the one hand, if $T = M(T)$ has unit diagonal then, from Lemma 3.3,

$$\operatorname{cond}(T, e) = \operatorname{cond}(T) = \|2T^{-1} - I\|_\infty \approx 2\frac{\kappa_\infty(T)}{\|T\|_\infty}.$$

This means, for example, that the system $Ux = b$, where $x = e$ and $U = T(1)$ in (3.4), is about as ill-conditioned with respect to componentwise relative perturbations in U as it is with respect to unstructured perturbations in U.

On the other hand, a triangular M-matrix system with a nonnegative right-hand side is very well-conditioned with respect to componentwise relative perturbations, irrespective of the size of κ.

LEMMA 3.4. *Suppose* $T = M(T)$ *and* $Tx = b \geqq 0$. *Then* $|T^{-1}||T||x| \leqq (2n-1)|x|$, *and hence* $\operatorname{cond}(T, x) \leqq 2n - 1$.

Proof. Write $T = D - U$, where $D = \operatorname{diag}(t_{ii}) \geqq 0$ and $U \geqq 0$ is strictly upper triangular. Then, using $(D^{-1}U)^n = 0$,

$$\begin{aligned} |T^{-1}||T| &= (I - D^{-1}U)^{-1}D^{-1} \cdot (D + U) \\ &= \sum_{i=0}^{n-1} (D^{-1}U)^i \cdot (I + D^{-1}U) \\ &= \sum_{i=0}^{n-1} (D^{-1}U)^i + \sum_{i=1}^{n-1} (D^{-1}U)^i. \end{aligned}$$

Now $0 \leqq b = Tx = (D - U)x$, so $Dx \geqq Ux$, that is, $x \geqq D^{-1}Ux$. Hence

$$|T^{-1}||T|x = \sum_{i=0}^{n-1} (D^{-1}U)^i x + \sum_{i=1}^{n-1} (D^{-1}U)^i x \leqq (n + n - 1)x,$$

which gives the result, since $x = T^{-1}b \geqq 0$. □

From Lemma 3.4 we obtain a result similar to Wilkinson's result (5) in § 2.

THEOREM 3.5. *The computed solution to the triangular system* $Tx = b$, *where* $T = M(T)$ *and* $b \geqq 0$, *satisfies*

$$|x - \hat{x}| \leqq d_n u|x| + O(u^2),$$

where $d_n = (2n-1)c_n$.

Proof. From (3.1) we have

$$\hat{x} = (T + E)^{-1}b = (T^{-1} - T^{-1}ET^{-1} + O(u^2))b,$$

and thus

$$|x - \hat{x}| \leqq c_n u|T^{-1}||T||x| + O(u^2).$$

The result follows from Lemma 3.4. □

Triangular systems of the type in Theorem 3.5 arise in computing estimates of $\|A^{-1}\|_\infty$ and of the smallest singular value of A, for triangular, bidiagonal, and

tridiagonal A [3], [12], [13]. They also occur in solving linear equations obtained from discretisation of certain elliptic partial differential equations, such as the Poisson equation on a rectangle, with zero boundary conditions and a positive forcing function (these problems yield symmetric positive definite M-matrices, and the LU factors of an M-matrix are themselves M-matrices).

4. Numerical experiments. We have carried out a variety of numerical experiments to confirm the analysis of § 3 and to gain further insight into the practical behaviour of the forward error in solution of a triangular system.

The computations were performed in PC-Matlab [16], which uses IEEE standard double-precision arithmetic (unit roundoff $2^{-52} \approx 2.2 \times 10^{-16}$). Three implementations of the back and forward substitution algorithms were tested: the vector sum version, and the inner product version with two different orderings of the terms in the inner product. In terms of back substitution for solving an $n \times n$ upper triangular system $Tx = b$, the three implementations are defined as follows.

```
VS:  x = b
     for i = n, n-1, ..., 1
          x_i = x_i / t_ii
          for j = 1, ..., i-1
               x_j = x_j - t_ji * x_i
          endfor
     endfor

IP1: for i = n, n-1, ..., 1
          s = 0
          for j = i+1, ..., n
               s = s + t_ij * x_j
          endfor
          x_i = (b_i - s) / t_ii
     endfor

IP2: for i = n, n-1, ..., 1
          s = 0
          for j = n, n-1, ..., i+1
               s = s + t_ij * x_j
          endfor
          x_i = (b_i - s) / t_ii
     endfor
```

(Each of these implementations is permissible for the triangular equation solver in the Level 2 BLAS [6].)

Various types of triangular matrices were generated. For each T seven linear systems $Tx = b$ were solved, defined by

$$\begin{aligned} &x_i \in N(0,1), && x = e, \\ &x_i = \alpha^{i-1}, && x_i = \alpha^{n-i}, \\ &b_i \in N(0,1), && b = e, \\ &b = \begin{cases} e_n & \text{if } T \text{ is upper triangular,} \\ e_1 & \text{if } T \text{ is lower triangular,} \end{cases} \end{aligned} \tag{4.1}$$

where $\alpha = 10^{-5/(n-1)} < 1$ and $N(0,1)$ denotes the normal distribution on $[0,1]$. Each

triangular system was solved four times: once in double precision using VS, and three times in single precision using VS, IP1, and IP2. Since PC-Matlab does not support single-precision arithmetic we simulated it by rounding the result of every arithmetic operation to 23 significant bits; this gives an effective unit roundoff $u = 2^{-23} \approx 1.2 \times 10^{-7}$. Each triangular system was generated in double precision and rounded to single precision before being solved.

For each single-precision solution $\hat{x}$ we computed the scaled componentwise and normwise relative errors

$$E_c = \max_{x_i \neq 0} \frac{|x_i - \hat{x}_i|}{u|x_i|}, \qquad E_n = \frac{\|x - \hat{x}\|_\infty}{u\|x\|_\infty}, \tag{4.2}$$

where for x we took the double-precision solution. Note that the relative errors are divided by u; thus values of order 1 for $E_c(E_n)$ correspond to all components of $\hat{x}$ (the largest component of $\hat{x}$) having all significant digits correct.

In our experiments very many of the triangular systems were indeed solved to high accuracy, E_c and E_n frequently being less than 100. The selected results reported below are not necessarily typical; although they do contain examples of high-accuracy solutions, they have been carefully chosen to illustrate extremes of numerical behaviour.

First we consider in detail an interesting example from [22, p. 233]. R is the Cholesky factor of $1.8144 \times A$, where A is the $(1{:}5, 2{:}6)$ submatrix of the Hilbert matrix. To the significant figures quoted,

$$R = \begin{bmatrix} 9.5247\,\text{E}-1 & 6.3498\,\text{E}-1 & 4.7624\,\text{E}-1 & 3.8099\,\text{E}-1 & 3.1749\,\text{E}-1 \\ 0 & 2.2450\,\text{E}-1 & 2.6940\,\text{E}-1 & 2.6940\,\text{E}-1 & 2.5657\,\text{E}-1 \\ 0 & 0 & 5.4991\,\text{E}-2 & 9.4270\,\text{E}-2 & 1.1784\,\text{E}-1 \\ 0 & 0 & 0 & 1.3607\,\text{E}-2 & 3.0237\,\text{E}-2 \\ 0 & 0 & 0 & 0 & 3.3806\,\text{E}-3 \end{bmatrix},$$

$$\text{cond}\,(R) = 1.36\,\text{E}1, \qquad \kappa_\infty(R) = 2.02\,\text{E}3,$$

$$\text{cond}\,(R^T) = 1.24\,\text{E}3, \qquad \kappa_\infty(R^T) = 1.52\,\text{E}3.$$

R does not satisfy (3.5) and is not an M-matrix. The values for cond show that the system $Rx = b$ is always well-conditioned with respect to componentwise relative perturbations in R, whereas $R^T x = b$ is, for some b, moderately ill-conditioned in the same sense. (The contrast in the conditioning in this example is noted also in [7, p. 264] and [22, p. 250].) Selected numerical results are displayed in Tables 4.1 and 4.2.

We offer the following comments on the results.

(1) For each solution method the normwise error is mostly predicted correctly, to within about an order of magnitude, by (3.2), and is significantly overestimated in the case of $Rx = b$ by (2.3).

TABLE 4.1
$Rx = b$.

	cond (R, x)	E_c(VS)	E_n(VS)	E_c(IP1)	E_n(IP1)	E_c(IP2)	E_n(IP2)
$x_i = \alpha^{n-i}$	7.82 E0	1.75 E3	3.12 E−2	6.87 E4	7.91 E−1	8.92 E4	1.44 E0
$x_i = \alpha^{i-1}$	1.08 E0	1.06 E0	7.93 E−2	1.06 E0	7.93 E−2	1.06 E0	7.93 E−2
$x = e$	1.36 E1	3.45 E0	3.45 E0	4.44 E0	4.44 E0	4.44 E0	4.44 E0

TABLE 4.2
$R^T x = b$.

	cond (R^T, x)	E_c(VS)	E_n(VS)	E_c(IP1)	E_n(IP1)	E_c(IP2)	E_n(IP2)
$x_i = \alpha^{n-i}$	2.48 E0	2.00 E0	2.44 E−1	5.44 E−1	2.44 E−1	2.00 E0	7.56 E−1
$x_i = \alpha^{i-1}$	7.04 E2	3.94 E5	1.25 E0	3.84 E7	1.21 E2	3.84 E7	1.21 E2
$x_i \in N(0, 1)$	3.83 E2	4.86 E1	4.86 E1	1.44 E1	1.44 E1	1.05 E2	1.05 E2
$x = e$	1.24 E3	6.69 E1	6.69 E1	1.55 E2	1.55 E2	1.41 E2	1.41 E2

(2) The errors E_n(VS), E_n(IP1), and E_n(IP2) vary significantly relative to one another. For example, in Table 4.2, $E_n(\text{IP1}) \approx 100\, E_n(\text{VS})$ for the second x, while $E_n(\text{IP1}) \approx \frac{1}{3} E_n(\text{VS})$ for the third x.

(3) The results for $x_i = \alpha^{n-i}$ in Table 4.1 and for $x_i = \alpha^{i-1}$ in Table 4.2 indicate that "graded" solution vectors x, whose components decay steadily in absolute value in the order in which they are computed, are a bad case as regards obtaining high componentwise relative accuracy. Intuitively this is to be expected, since the small components are obtained as linear combinations of the larger ones, and so severe cancellation is likely to occur.

Next we consider Cholesky factors of the Pascal matrix P_{15}, where the symmetric positive definite $n \times n$ matrix $P_n = (p_{ij})$ is defined by $p_{i1} = p_{1i} \equiv 1$, $p_{ij} = p_{i,j-1} + p_{i-1,j} (i, j > 1)$. Let U be the Cholesky factor without pivoting ($P_{15} = U^T U$) and let U_p be the Cholesky factor with complete pivoting ($\Pi^T P_{15} \Pi = U_p^T U_p$). We have

$$\begin{aligned} &\text{cond}\,(U) = 1.58\ \text{E6}, && \kappa_\infty(U) = 4.14\ \text{E7}, \\ &\text{cond}\,(U_p) = 2.25\ \text{E1}, && \kappa_\infty(U_p) = 5.13\ \text{E7}, \\ &\text{cond}\,(M(U)) = 2.24\ \text{E13}, && \kappa_\infty(M(U)) = 7.21\ \text{E16}, \\ &\text{cond}\,(M(U_p)) = 9.47\ \text{E1}, && \kappa_\infty(M(U_p)) = 8.47\ \text{E8}. \end{aligned}$$

Note the dramatic reduction in cond (U) brought about by pivoting (Lemma 3.1 yields the bound cond $(U_p) \leqq 2^{15} - 1 = 32767$); $\kappa_\infty(U)$ is almost unchanged, since $\kappa_2(U) = \kappa_2(U_p) = \kappa_2(P_{15})^{1/2}$.

Selected results are given in Tables 4.3–4.5. Notable features are as follows.

- Table 4.3 illustrates how the conditioning, and the achieved accuracy, can vary greatly with the right-hand side.

TABLE 4.3
$Ux = b$.

	cond (U, x)	E_c(VS)	E_n(VS)	E_c(IP1)	E_n(IP1)	E_c(IP2)	E_n(IP2)
$x_i = \alpha^{i-1}$	3.93 E1	4.10 E2	8.92 E0	8.93 E2	5.31 E0	1.70 E3	1.13 E1
$x = e$	1.49 E6	2.02 E4	1.82 E4	1.04 E5	9.31 E4	4.74 E5	4.30 E5

TABLE 4.4
$U_p x = b$.

	cond (U_p, x)	E_c(VS)	E_n(VS)	E_c(IP1)	E_n(IP1)	E_c(IP2)	E_n(IP2)
$b = e$	1.11 E1	1.19 E4	4.83 E0	4.86 E3	8.72 E−1	1.19 E4	4.83 E0
$x = e$	2.25 E0	9.40 E0	9.40 E0	5.85 E0	5.85 E0	6.27 E0	6.27 E0

TABLE 4.5
$M(U)x = b$.

	cond $(M(U), x)$	E_c(VS)	E_n(VS)	E_c(IP1)	E_n(IP1)	E_c(IP2)	E_n(IP2)
$x_i = \alpha^{n-i}$	4.82 E7	4.70 E4	3.05 E4	3.24 E6	2.08 E6	3.84 E6	2.48 E6
$b = e$	2.16 E1	2.21 E0	2.21 E0	1.99 E0	1.42 E0	2.21 E0	2.21 E0
$b = e_n$	2.16 E0	1.38 E0	9.35 E−1	4.17 E0	2.35 E0	1.38 E0	9.35 E−1

- Table 4.4 shows that the componentwise relative error can be quite large even when the upper triangular matrix T satisfies (3.5) and cond (T, x) is small.
- The entries for $b = e$ and $x = e$ in Table 4.5 illustrate well the behaviour predicted by Lemma 3.4 and Theorem 3.5.

Finally we report on an experiment in which random 10×10 symmetric positive definite matrices $A = V^T \Lambda V$ were formed, where V denotes a random orthogonal matrix constructed by the method of [19], and $\Lambda = \text{diag}(\lambda_i)$ with the eigenvalues from the exponential distribution $\lambda_i = \beta^i$, or the sharp-break distribution $\lambda_1 = \cdots \lambda_9 = 1$, $\lambda_{10} = \beta$. The Cholesky decomposition of A was computed both with and without pivoting, yielding triangular factors G and G_p, respectively. Throughout, β was chosen so that $\kappa_2(A) = 10^{12}$, and thus $\kappa_2(G) = \kappa_2(G_p) = 10^6$. The system $G^T G x = b$, that is,

$$G^T y = b, \qquad Gx = y,$$

was solved for the first six x and b in (4.1) together with $b = G^T e$ (and similarly for G_p). A different random A was used for each different x and b. We report results for the vector sum algorithm only. Three errors are of interest: those in the forward substitution $G^{-T} b - \hat{y}$ and the back substitution $G^{-1}\hat{y} - \hat{x}$, and the overall error $x - \hat{x}$. We report the normwise errors, denoted $E_n(G^T)$, $E_n(G)$, and $E_n(x)$, respectively (and defined as in (4.2)). The results, summarized in Tables 4.6–4.9, display several interesting features.

- The accuracy of the solution to the *coupled* triangular systems $G^T G x = b$ (and $G_p^T G_p x = b$) depends very much on the right-hand side, as the $E_n(x)$ values show.

TABLE 4.6
$G^T G x = b$. *Exponential* λ_i *distribution.*

	cond (G^T)	cond (G^T, y)	$E_n(G^T)$	cond (G)	cond (G, x)	$E_n(G)$	$E_n(x)$
$b_i \in N(0,1)$	8.35 E5	2.21 E2	1.50 E1	8.04 E1	4.54 E1	3.26 E0	1.44 E1
$x_i \in N(0,1)$	5.20 E5	3.47 E5	1.45 E4	2.17 E2	3.94 E1	4.37 E0	4.28 E5
$b = G^T e$	7.60 E5	7.57 E5	1.39 E4	9.33 E1	2.78 E1	2.28 E0	3.97 E3

TABLE 4.7
$G_p^T G_p x = b$. *Exponential* λ_i *distribution.*

	cond (G_p^T)	cond (G_p^T, y)	$E_n(G_p^T)$	cond (G_p)	cond (G_p, x)	$E_n(G_p)$	$E_n(x)$
$b_i \in N(0,1)$	7.90 E5	1.25 E1	3.88 E0	1.17 E1	5.37 E0	8.11 E−1	3.03 E0
$x_i \in N(0,1)$	6.05 E5	1.74 E0	1.61 E0	1.05 E1	3.05 E0	5.75 E−1	1.64 E0
$b = G_p^T e$	9.50 E5	9.41 E5	3.99 E4	1.64 E1	7.67 E0	1.24 E0	3.99 E4

TABLE 4.8
$G^T Gx = b$. *Sharp break* λ_i *distribution.*

	cond (G^T)	cond (G^T, y)	$E_n(G^T)$	cond (G)	cond (G, x)	$E_n(G)$	$E_n(x)$
$b_i \in N(0,1)$	3.55 E6	1.48 E1	1.37 E0	3.71 E0	2.94 E0	5.93 E−1	1.85 E0
$x_i \in N(0,1)$	2.48 E6	6.98 E5	1.00 E5	4.55 E1	6.11 E0	6.93 E−1	8.39 E6
$b = G^T e$	3.64 E6	3.64 E6	7.67 E5	4.13 E1	9.42 E0	2.84 E0	7.68 E5

TABLE 4.9
$G_p^T G_p x = b$. *Sharp break* λ_i *distribution.*

	cond (G_p^T)	cond (G_p^T, y)	$E_n(G_p^T)$	cond (G_p)	cond (G_p, x)	$E_n(G_p)$	$E_n(x)$
$b_i \in N(0,1)$	3.55 E6	1.62 E1	3.78 E0	3.48 E0	2.90 E0	2.58 E−1	3.88 E0
$x_i \in N(0,1)$	2.38 E6	3.91 E0	1.87 E0	2.03 E0	1.54 E0	1.04 E0	2.62 E0
$b = G_p^T e$	3.54 E6	3.26 E6	1.09 E6	3.99 E0	3.99 E0	9.93 E−1	1.09 E6

• In all cases, cond (G), cond $(G_p) \ll \kappa_\infty(G) \approx \kappa_\infty(G_p) \approx 10^6$, emphasising once again the possible disparity between the values of the condition numbers cond and κ_∞.

• cond $(G^T) \gg$ cond (G) throughout (and similarly for G_p), showing even more strikingly than in the Hilbert matrix test that a triangular matrix may be much more ill-conditioned than its transpose. This contrast in the condition numbers is reflected in the forward errors: in about half the examples quoted, the error in the forward substitution greatly exceeds that in the back substitution!

Concerning the previous comment, an interesting heuristic emerged during our experiments: in practice the upper triangular matrices T arising in LU (by Gaussian elimination), QR, and Cholesky decompositions—all either with or without pivoting—tend to satisfy $\kappa_\infty(T) \approx$ cond $(T^T) >$ cond (T).

A partial explanation can be given in terms of the scaling properties of cond, together with another heuristic, namely that any small diagonal elements of T tend to appear towards the (n, n) position. It is tantalising to ask what is the implication of this heuristic for the choice of normalisation in a decomposition. For example, in the last experiment, can we avoid the effects of a large cond (G^T) by employing instead an LDL^T decomposition (L unit lower triangular), where, cond (L) and cond (L^T) will certainly be quite small? The answer is no, as the following short analysis shows.

Consider a system $LDUx = b$, where L is lower triangular, D is diagonal, and U is upper triangular. From Theorem 2.1 the computed solution $\hat{x}$ satisfies

$$(L+\Delta L)(D+\Delta D)(U+\Delta U)\hat{x} = b, \tag{4.3}$$

$$|\Delta X| \leqq c_n u |X|, \qquad X = L, D, U. \tag{4.4}$$

After some manipulation of (4.3) we obtain

$$x - \hat{x} = (U^{-1}D^{-1}L^{-1}\Delta LDU + U^{-1}D^{-1}\Delta DU + U^{-1}\Delta U)x + O(u^2),$$

which gives, using (4.4) and $|D^{-1}||D| = I$,

$$|\hat{x} - x| \leqq c_n u(|U^{-1}||D^{-1}||L^{-1}||L|D||U| + 2|U^{-1}||U|)|x| + O(u^2).$$

This bound indicates that any attempt to manipulate the normalisation to the advantage of the forward error will, in general, be futile. For whether D is combined with L or with U (e.g., Crout reduction or Gaussian elimination), or left separate, the dominant term in the bound is unchanged.

Finally we mention two practical issues. First, for $n \times n$ triangular T the condition number cond (T, x) can be estimated in $O(n^2)$ operations, without computing T^{-1},

using an algorithm from [9], [14] (see [1] for the details). We stress, however, that in most applications cond (T, x) will not be of direct interest; rather, some condition number for the overall problem (e.g., $\kappa(A)$ or cond (A, x) for $Ax = b$) will be the most appropriate quantity to examine (see [1], [13]). Second, since Theorem 2.1 shows that the backward error in solving a triangular system is about as small as we could reasonably expect, it is not worth doing iterative refinement in single precision for triangular systems. In the LAPACK project, single-precision iterative refinement routines are being supplied for all linear systems except triangular ones [2].

5. Conclusions. Triangular systems are *usually* solved to high accuracy, but various contrary types of behaviour are possible within the freedom afforded by the bounds of § 3. The condition number cond (T, x) is the key to understanding and predicting the behaviour of the forward error in the solution of a triangular system $Tx = b$. Some of the phenomena we have identified are as follows.

(1) The computed solution $\hat{x}$ to $Tx = b$ may be highly accurate irrespective of the size of $\kappa_\infty(T)$.

(2) $\hat{x}$ may have a small normwise relative error but a large componentwise relative error, but in our experience this is uncommon.

(3) The accuracy of $\hat{x}$ may depend very much on the right-hand side b (if cond (T) is large). (Note that Skeel draws this conclusion in [17] for general linear systems.)

(4) T^T may be much more ill-conditioned than T, i.e., cond $(T^T) \gg$ cond (T).

Points 3 and 4 do not seem to be well known. A likely reason is that in practice such subtle behaviour is masked by the algorithm that leads to the triangular system. For example, rounding errors in the reduction phase of Gaussian elimination tend to dominate those for the triangular solves.

Our experiments have shown that the forward errors can vary significantly with the implementation of a substitution algorithm: merely reordering the inner products can change the forward error by orders of magnitude. This is not at all surprising, since it is well known that the forward error in a summation can be very sensitive to the order of summation, but the fact is easily overlooked. An important implication is that we must take extreme care when basing judgments of competing algorithms on comparisons of their forward errors!

Our work provides some additional support for the use of pivoting in QR and Cholesky decompositions. The value of cond (T) is usually smaller if pivoting is used than if it is not, and consequently triangular systems may be solved more accurately (see Lemma 3.1, Theorem 3.2, and Tables 4.3–4.4). This fact may help to explain the empirical observation that least-squares problems tend to be solved to higher accuracy when pivoting is used in the QR decomposition [15].

Acknowledgments. I thank the referees for carefully reading the manuscript and offering helpful suggestions.

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] C. H. BISCHOF, J. W. DEMMEL, J. J. DONGARRA, J. J. DU CROZ, A. GREENBAUM, S. J. HAMMARLING, AND D. C. SORENSEN, *Provisional contents*, LAPACK Working Note 5, Report ANL-88-38, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1988.

[3] J. W. DEMMEL AND W. KAHAN, *Computing small singular values of bidiagonal matrices with guaranteed high relative accuracy*, LAPACK Working Note 3, Tech. Memorandum 110, Argonne National Laboratory, Argonne, IL, 1988; SIAM J. Sci. Statist. Comput., to appear.

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

[5] J. J. DONGARRA, J. J. DU CROZ, S. J. HAMMARLING, AND R. J. HANSON, *An extended set of Fortran basic linear algebra subprograms*, ACM Trans. Math. Software, 14 (1988), pp. 1–17.

[6] ———, *Algorithm* 656: *An extended set of Fortran basic linear algebra subprograms: Model implementation and test programs*, ACM Trans. Math. Software, 14 (1988), pp. 18–32.

[7] R. FLETCHER, *Expected conditioning*, IMA J. Numer. Anal., 5 (1985), pp. 247–273.

[8] W. GOVAERTS AND J. D. PRYCE, *Block elimination with one iterative refinement solves bordered linear systems accurately*, Report AM-88-01, School of Mathematics, University of Bristol, Bristol, U.K., 1988.

[9] W. W. HAGER, *Condition estimates*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 311–316.

[10] M. T. HEATH AND C. H. ROMINE, *Parallel solution of triangular systems on distributed-memory multiprocessors*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 558–588.

[11] D. HELLER, *A survey of parallel algorithms in numerical linear algebra*, SIAM Rev., 20 (1978), pp. 740–777.

[12] N. J. HIGHAM, *Efficient algorithms for computing the condition number of a tridiagonal matrix*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 150–165.

[13] ———, *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29 (1987), pp. 575–596.

[14] ———, *Fortran codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation*, ACM Trans. Math. Software, 14 (1988), pp. 381–396.

[15] L. S. JENNINGS AND M. R. OSBORNE, *A direct error analysis for least squares*, Numer. Math., 22 (1974), pp. 325–332.

[16] C. B. MOLER, J. N. LITTLE, AND S. BANGERT, PC-*Matlab User's Guide*, The Math Works, Inc., Sherborn, MA, 1987.

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

[18] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[19] ———, *The efficient generation of random orthogonal matrices with an application to condition estimators*, SIAM J. Numer. Anal., 17 (1980), pp. 403–409.

[20] J. H. WILKINSON, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281–330.

[21] ———, *Rounding Errors in Algebraic Processes*, *Notes on Applied Science*, No. 32, Her Majesty's Stationery Office, London, 1963.

[22] ———, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.