Triangular systems play a fundamental role in matrix computations. Many methods are built on the idea of reducing a problem to the solution of one or more triangular systems, including virtually all direct methods for solving linear systems. On serial computers triangular systems are universally solved by the standard back and forward substitution algorithms. For parallel computation there are several alternative methods, one of which we analyse in §8.4.

Backward error analysis for the substitution algorithms is straightforward and the conclusion is well known: the algorithms are extremely stable. The behaviour of the forward error, however, is intriguing, because the forward error is often surprisingly small—much smaller than we would predict from the normwise condition number κ, or, sometimes, even the componentwise condition number cond. The quotes from Stewart and Wilkinson at the start of this chapter emphasize the high accuracy that is frequently observed in practice. The analysis we give in this chapter provides a partial explanation for the observed accuracy of the substitution algorithms. In particular, it reveals three important but nonobvious properties:

- the accuracy of the computed solution from substitution depends strongly on the right-hand side;
- a triangular matrix may be much more or less ill conditioned than its transpose; and
- the use of pivoting in LU, QR, and Cholesky factorizations can greatly improve the conditioning of a resulting triangular system.

As well as deriving backward and forward error bounds, we show how to compute upper and lower bounds for the inverse of a triangular matrix.

8.1. Backward Error Analysis

Recall that for an upper triangular matrix $U \in \mathbb{R}^{n\times n}$ the system $Ux = b$ can be solved using the formula $x_i = (b_i - \sum_{j=i+1}^{n} u_{ij}x_j)/u_{ii}$, which yields the components of x in order from last to first.

Algorithm 8.1 (back substitution). Given a nonsingular upper triangular matrix $U \in \mathbb{R}^{n\times n}$ this algorithm solves the system $Ux = b$.

```
x_n = b_n/u_nn
for i = n-1: -1: 1
    s = b_i
    for j = i+1: n
        s = s - u_ij x_j
    end
    x_i = s/u_ii
end
```


We will not state the analogous algorithm for solving a lower triangular system, forward substitution. All the results below for back substitution have obvious analogues for forward substitution. Throughout this chapter T denotes a matrix that can be upper or lower triangular.

To analyse the errors in substitution we need the following lemma.

Lemma 8.2. *Let $y = (c - \sum_{i=1}^{k-1} a_i b_i)/b_k$ be evaluated in floating point arithmetic according to*

```
s = c
for i = 1: k-1
    s = s - a_i b_i
end
y = s/b_k
```

Then the computed $\widehat{y}$ satisfies

$$b_k\widehat{y}(1+\theta_k) = c - \sum_{i=1}^{k-1} a_i b_i (1+\theta_i), \tag{8.1}$$

where $|\theta_i| \le \gamma_i = iu/(1-iu)$.

Proof. Analysis very similar to that leading to (3.2) shows that $\widehat{s} := fl(c - \sum_{i=1}^{k-1} a_i b_i)$ satisfies

$$\widehat{s} = c(1+\delta_1)\dots(1+\delta_{k-1}) - \sum_{i=1}^{k-1} a_i b_i (1+\epsilon_i)(1+\delta_i)\dots(1+\delta_{k-1}),$$

where $|\epsilon_i|, |\delta_i| \le u$. The final division yields, using (2.5), $\widehat{y} = fl(\widehat{s}/b_k) = \widehat{s}/(b_k(1+\delta_k))$, $|\delta_k| \le u$, so that, after dividing through by $(1+\delta_1)\dots(1+\delta_{k-1})$, we have

$$b_k\widehat{y}\frac{1+\delta_k}{(1+\delta_1)\dots(1+\delta_{k-1})} = c - \sum_{i=1}^{k-1} a_i b_i \frac{1+\epsilon_i}{(1+\delta_1)\dots(1+\delta_{i-1})}.$$

The result is obtained on invoking Lemma 3.1. □

Two remarks are in order. First, we chose the particular form of (8.1), in which c is not perturbed, in order to obtain a backward error result for $Ux = b$ in which b is not perturbed. Second, we carefully kept track of the terms $1+\delta_i$ in the proof, so as to obtain the best possible constants. Direct application of the lemma to Algorithm 8.1 yields a backward error result.

Theorem 8.3. *The computed solution $\widehat{x}$ from Algorithm 8.1 satisfies*

$$(U + \Delta U)\widehat{x} = b, \qquad |\Delta u_{ij}| \le \begin{cases} \gamma_{n-i+1}|u_{ii}|, & i = j, \\ \gamma_{|i-j|}|u_{ij}|, & i \ne j. \end{cases} \qquad \square$$

Theorem 8.3 holds only for the particular ordering of arithmetic operations used in Algorithm 8.1. A result that holds for any ordering is a consequence of the next lemma.

Lemma 8.4. *If $y = (c - \sum_{i=1}^{k-1} a_i b_i)/b_k$ is evaluated in floating point arithmetic, then, no matter what the order of evaluation,*

$$b_k\widehat{y}(1 + \theta_k^{(0)}) = c - \sum_{i=1}^{k-1} a_i b_i (1 + \theta_k^{(i)}),$$

where $|\theta_k^{(i)}| \le \gamma_k$ for all i. If $b_k = 1$, so that there is no division, then $|\theta_k^{(i)}| \le \gamma_{k-1}$ for all i.

Proof. The result is not hard to see after a little thought, but a formal proof is tedious to write down. Note that the ordering used in Lemma 8.2 is the one for which this lemma is least obvious! The last part of the lemma is useful when analysing unit lower triangular systems, and in various other contexts. $\square$

Theorem 8.5. *Let the triangular system $Tx = b$, where $T \in \mathbb{R}^{n\times n}$ is nonsingular, be solved by substitution, with any ordering. Then the computed solution $\widehat{x}$ satisfies*

$$(T + \Delta T)\widehat{x} = b, \qquad |\Delta T| \le \gamma_n |T|. \qquad \square$$

In technical terms, this result says that $\widehat{x}$ has a tiny componentwise relative backward error. In other words, the backward error is about as small as we could possibly hope.

In most of the remaining error analyses in this book, we will derive results that, like the one in Theorem 8.5, do not depend on the ordering of the arithmetic operations. Results of this type are more general, usually no less informative, and easier to derive, than ones that depend on the ordering. However, it is important to realise that *the actual error does depend on the ordering*, possibly strongly so for certain data. This point is clear from Chapter 4 on summation.

8.2. Forward Error Analysis

From Theorems 8.5 and 7.4 there follows the forward error bound

$$\frac{\|x - \widehat{x}\|_\infty}{\|x\|_\infty} \le \frac{\operatorname{cond}(T,x)\gamma_n}{1 - \operatorname{cond}(T)\gamma_n},$$

where

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

This bound can, of course, be arbitrarily smaller than the corresponding bound involving $\kappa_\infty(T) = \|T\|_\infty \|T^{-1}\|_\infty$, for the reasons explained in Chapter 7. For further insight, note that, 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 which are large relative to the diagonal elements. Significantly, because of its row scaling invariance, $\operatorname{cond}(T,x)$ is susceptible only to the second source.

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

$$U(\alpha) = (u_{ij}), \qquad u_{ij} = \begin{cases} 1, & i = j, \\ -\alpha, & i < j, \end{cases} \tag{8.2}$$

for which

$$\left(U(\alpha)^{-1}\right)_{ij} = \begin{cases} 1, & i = j, \\ \alpha(1+\alpha)^{j-i-1}, & j > i. \end{cases} \tag{8.3}$$

We have $\operatorname{cond}(U(\alpha), e) = \operatorname{cond}(U(\alpha)) \sim 2\alpha^{n-1}$ 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 $\operatorname{cond}(T,x) = 1$, and the solution comprises the computation of just a single scalar reciprocal.

To gain further insight we consider special classes of triangular matrices, beginning with one produced by certain standard factorizations with pivoting. In all the results below, the triangular matrices are assumed to be $n \times n$ and nonsingular, and $\widehat{x}$ is the computed solution from substitution.

Lemma 8.6. *Suppose the upper triangular matrix $U \in \mathbb{R}^{n\times n}$ satisfies*

$$|u_{ii}| \ge |u_{ij}| \quad \text{for all } j > i. \tag{8.4}$$

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

Proof. We can write $W = |V^{-1}||V|$ where $V = D^{-1}U$ and $D = \mathrm{diag}(u_{ii})$. The matrix V is unit upper triangular with $|v_{ij}| \le 1$, and it is easy to show that $|(V^{-1})_{ij}| \le 2^{j-i-1}$ for $j > i$. Thus, for $j > i$,

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

Theorem 8.7. *Under the conditions of Lemma 8.6, the computed solution $\widehat{x}$ to $Ux = b$ obtained by substitution satisfies*

$$|x_i - \widehat{x}_i| \le 2^{n-i+1}\gamma_n \max_{j \ge i} |\widehat{x}_j|, \qquad i = 1\colon n.$$

Proof. From Theorem 8.5 we have

$$|x - \widehat{x}| = |U^{-1}\Delta U \widehat{x}| \le \gamma_n |U^{-1}||U||\widehat{x}|.$$

Using Lemma 8.6 we obtain

$$|x_i - \widehat{x}_i| \le \gamma_n \sum_{j=i}^{n} w_{ij}|\widehat{x}_j| \le \gamma_n \max_{j \ge i} |\widehat{x}_j| \sum_{j=i}^{n} 2^{j-i} \le 2^{n-i+1}\gamma_n \max_{j \ge i} |\widehat{x}_j|. \quad \square$$

Lemma 8.6 shows that for matrices satisfying (8.4), $\mathrm{cond}(T)$ is bounded for fixed n, no matter how large $\kappa(T)$. The bounds for $|x_i - \widehat{x}_i|$ in Theorem 8.7, 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 8.6 and Theorem 8.7 hold for lower triangular L satisfying

$$|l_{ii}| \ge |l_{ij}| \quad \text{for all } j < i. \tag{8.5}$$

Note, however, that if the upper triangular matrix T satisfies (8.4) then T^T does not necessarily satisfy (8.5). In fact, $\mathrm{cond}(T^T)$ can be arbitrarily large, as shown by the example

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

$$\mathrm{cond}(T) = 5, \quad \mathrm{cond}(T^T) = 1 + \frac{2}{\epsilon}.$$

An important conclusion is that a triangular system $Tx = b$ can be much more or less ill conditioned than the system $T^Ty = c$, even if T satisfies (8.4).

Theorem 8.7, 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 factorizations with complete pivoting and column pivoting, respectively.

Next, we consider triangular T satisfying

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

It is easy to see that such a matrix has an inverse with nonnegative elements, and hence is an M-matrix (for definitions of an M-matrix see Appendix B). Associated with any square matrix A is the *comparison matrix*:

$$M(A) = (m_{ij}), \qquad m_{ij} = \begin{cases} |a_{ii}|, & i = j, \\ -|a_{ij}|, & i \ne j. \end{cases} \tag{8.6}$$

For any nonsingular triangular T, $M(T)$ is an M-matrix. Furthermore, it is easy to show that $|T^{-1}| \le M(T)^{-1}$ (see Theorem 8.11).

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

Lemma 8.8. *For any triangular T,*

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

Proof. The inequality follows from $|T^{-1}| \le M(T)^{-1}$, together with $|T| = |M(T)|$. Since $M(T)^{-1} \ge 0$, we have

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

which yields the equality. $\square$

If $T = M(T)$ has unit diagonal then, using Lemma 8.8,

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

This means, for example, that the system $U(1)x = b$ (see (8.2)), where $x = e$, is about as ill conditioned with respect to componentwise relative perturbations in $U(1)$ as it is with respect to normwise perturbations in $U(1)$.