2

Linear Equation Solving

2.1. Introduction

This chapter discusses perturbation theory, algorithms, and error analysis for solving the linear equation $Ax = b$. The algorithms are all variations on Gaussian elimination. They are called *direct methods*, because in the absence of roundoff error they would give the exact solution of $Ax = b$ after a finite number of steps. In contrast, Chapter 6 discusses *iterative methods*, which compute a sequence $x_0, x_1, x_2, \ldots$ of ever better approximate solutions of $Ax = b$; one stops iterating (computing the next x_{i+1}) when x_i is accurate enough. Depending on the matrix A and the speed with which x_i converges to $x = A^{-1}b$, a direct method or an iterative method may be faster or more accurate. We will discuss the relative merits of direct and iterative methods at length in Chapter 6. For now, we will just say that direct methods are the methods of choice when the user has no special knowledge about the source[7] of matrix A or when a solution is required with guaranteed stability and in a guaranteed amount of time.

The rest of this chapter is organized as follows. Section 2.2 discusses perturbation theory for $Ax = b$; it forms the basis for the practical error bounds in section 2.4. Section 2.3 derives the Gaussian elimination algorithm for dense matrices. Section 2.4 analyzes the errors in Gaussian elimination and presents practical error bounds. Section 2.5 shows how to improve the accuracy of a solution computed by Gaussian elimination, using a simple and inexpensive iterative method. To get high speed from Gaussian elimination and other linear algebra algorithms on contemporary computers, care must be taken to organize the computation to respect the computer memory organization; this is discussed in section 2.6. Finally, section 2.7 discusses faster variations of Gaussian elimination for matrices with special properties commonly arising in practice, such as symmetry ($A = A^T$) or sparsity (when many entries of A are zero).

[7] For example, in Chapter 6 we consider the case when A arises from approximating the solution to a particular differential equation, Poisson's equation.

Sections 2.2.1 and 2.5.1 discuss recent innovations upon which the software in the LAPACK library depends.

There are a variety of open problems, which we shall mention as we go along.

2.2. Perturbation Theory

Suppose $Ax = b$ and $(A + \delta A)\hat{x} = b + \delta b$; our goal is to bound the norm of $\delta x \equiv \hat{x} - x$. Later, $\hat{x}$ will be the computed solution of $Ax = B$. We simply subtract these two equalities and solve for δx: one way to do this is to take

$$\begin{array}{rrcl} & (A+\delta A)(x+\delta x) & = & b+\delta b \\ - & [Ax & = & b] \\ \hline & \delta Ax + (A+\delta A)\delta x & = & \delta b \end{array}$$

and rearrange to get

$$\delta x = A^{-1}(-\delta A\hat{x} + \delta b). \tag{2.1}$$

Taking norms and using part 1 of Lemma 1.7 as well as the triangle inequality for vector norms, we get

$$\|\delta x\| \leq \|A^{-1}\|(\|\delta A\| \cdot \|\hat{x}\| + \|\delta b\|). \tag{2.2}$$

(We have assumed that the vector norm and matrix norm are consistent, as defined in section 1.7. For example, any vector norm and its induced matrix norm will do.) We can further rearrange this inequality to get

$$\frac{\|\delta x\|}{\|\hat{x}\|} \leq \|A^{-1}\| \cdot \|A\| \cdot \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \cdot \|\hat{x}\|} \right). \tag{2.3}$$

The quantity $\kappa(A) = \|A^{-1}\| \cdot \|A\|$ is the *condition number*[8] of the matrix A, because it measures the relative change $\frac{\|\delta x\|}{\|\hat{x}\|}$ in the answer as a multiple of the relative change $\frac{\|\delta A\|}{\|A\|}$ in the data. (To be rigorous, we need to show that inequality (2.2) is an equality for some nonzero choice of δA and δb; otherwise $\kappa(A)$ would only be an upper bound on the condition number. See Question 2.3.) The quantity multiplying $\kappa(A)$ will be small if δA and δb are small, yielding a small upper bound on the relative error $\frac{\|\delta x\|}{\|\hat{x}\|}$.

The upper bound depends on δx (via $\hat{x}$), which makes it seem hard to interpret, but it is actually quite useful in practice, since we know the computed solution $\hat{x}$ and so can straightforwardly evaluate the bound. We can also derive a theoretically more attractive bound that does not depend on δx as follows:

[8]More pedantically, it is the condition number with respect to the problem of matrix inversion. The problem of finding the eigenvalues of A, for example, has a different condition number.

LEMMA 2.1. *Let* $\|\cdot\|$ *satisfy* $\|AB\| \leq \|A\| \cdot \|B\|$. *Then* $\|X\| < 1$ *implies that* $I - X$ *is invertible,* $(I-X)^{-1} = \sum_{i=0}^{\infty} X^i$, *and* $\|(I-X)^{-1}\| \leq \frac{1}{1-\|X\|}$.

Proof. The sum $\sum_{i=0}^{\infty} X^i$ is said to converge if and only if it converges in each component. We use the fact (from applying Lemma 1.4 to Example 1.6) that for any norm, there is a constant c such that $|x_{jk}| \leq c \cdot \|X\|$. We then get $|(X^i)_{jk}| \leq c \cdot \|X^i\| \leq c \cdot \|X\|^i$, so each component of $\sum X^i$ is dominated by a convergent geometric series $\sum c\|X\|^i = \frac{c}{1-\|X\|}$ and must converge. Therefore $S_n = \sum_{i=0}^{n} X^i$ converges to some S as $n \to \infty$, and $(I-X)S_n = (I-X)(I + X + X^2 + \cdots + X^n) = I - X^{n+1} \to I$ as $n \to \infty$, since $\|X^i\| \leq \|X\|^i \to 0$. Therefore $(I-X)S = I$ and $S = (I-X)^{-1}$. The final bound is $\|(I-X)^{-1}\| = \|\sum_{i=0}^{\infty} X^i\| \leq \sum_{i=0}^{\infty} \|X^i\| \leq \sum_{i=0}^{\infty} \|X\|^i = \frac{1}{1-\|X\|}$. □

Solving our first equation $\delta A x + (A + \delta A)\delta x = \delta b$ for δx yields

$$\begin{aligned} \delta x &= (A+\delta A)^{-1}(-\delta A x + \delta b) \\ &= [A(I + A^{-1}\delta A)]^{-1}(-\delta A x + \delta b) \\ &= (I + A^{-1}\delta A)^{-1} A^{-1}(-\delta A x + \delta b). \end{aligned}$$

Taking norms, dividing both sides by $\|x\|$, using part 1 of Lemma 1.7 and the triangle inequality, and assuming that δA is small enough so that $\|A^{-1}\delta A\| \leq \|A^{-1}\| \cdot \|\delta A\| < 1$, we get the desired bound:

$$\begin{aligned} \frac{\|\delta x\|}{\|x\|} &\leq \|(I + A^{-1}\delta A)^{-1}\| \cdot \|A^{-1}\| \left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|} \right) \\ &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|\delta A\|} \left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|} \right) \quad \text{by Lemma 2.1} \\ &= \frac{\|A^{-1}\| \cdot \|A\|}{1 - \|A^{-1}\| \cdot \|A\| \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \cdot \|x\|} \right) \\ &\leq \frac{\kappa(A)}{1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right) \qquad (2.4) \\ &\quad \text{since } \|b\| = \|Ax\| \leq \|A\| \cdot \|x\|. \end{aligned}$$

This bound expresses the relative error $\frac{\|\delta x\|}{\|x\|}$ in the solution as a multiple of the relative errors $\frac{\|\delta A\|}{\|A\|}$ and $\frac{\|\delta b\|}{\|b\|}$ in the input. The multiplier, $\kappa(A)/(1 - \kappa(A)\frac{\|\delta A\|}{\|A\|})$, is close to the condition number $\kappa(A)$ if $\|\delta A\|$ is small enough.

The next theorem explains more about the assumption that $\|A^{-1}\| \cdot \|\delta A\| = \kappa(A) \cdot \frac{\|\delta A\|}{\|A\|} < 1$: it guarantees that $A + \delta A$ is nonsingular, which we need for δx to exist. It also establishes a geometric characterization of the condition number.

THEOREM 2.1. *Let* A *be nonsingular. Then*

$$\min \left\{ \frac{\|\delta A\|_2}{\|A\|_2} : A + \delta A \text{ singular} \right\} = \frac{1}{\|A^{-1}\|_2 \cdot \|A\|_2} = \frac{1}{\kappa(A)}.$$

Therefore, the distance to the nearest singular matrix (ill-posed problem) $= \frac{1}{\text{condition number}}$.

Proof. It is enough to show $\min \{\|\delta A\|_2 : A + \delta A \text{ singular}\} = \frac{1}{\|A^{-1}\|_2}$.

To show this minimum is at least $\frac{1}{\|A^{-1}\|_2}$, note that if $\|\delta A\|_2 < \frac{1}{\|A^{-1}\|_2}$, then $1 > \|\delta A\|_2 \cdot \|A^{-1}\|_2 \geq \|A^{-1}\delta A\|_2$, so Lemma 2.1 implies that $I + A^{-1}\delta A$ is invertible, and so $A + \delta A$ is invertible.

To show the minimum equals $\frac{1}{\|A^{-1}\|_2}$, we construct a δA of norm $\frac{1}{\|A^{-1}\|_2}$ such that $A + \delta A$ is singular. Note that since $\|A^{-1}\|_2 = \max_{x \neq 0} \frac{\|A^{-1}x\|_2}{\|x\|_2}$, there exists an x such that $\|x\|_2 = 1$ and $\|A^{-1}\|_2 = \|A^{-1}x\|_2 > 0$. Now let $y = \frac{A^{-1}x}{\|A^{-1}x\|_2} = \frac{A^{-1}x}{\|A^{-1}\|_2}$ so $\|y\|_2 = 1$. Let $\delta A = \frac{-xy^T}{\|A^{-1}\|_2}$.

Then

$$\|\delta A\|_2 = \max_{z \neq 0} \frac{\|xy^T z\|_2}{\|A^{-1}\|_2 \, \|z\|_2} = \max_{z \neq 0} \frac{|y^T z|}{\|z\|_2} \frac{\|x\|_2}{\|A^{-1}\|_2} = \frac{1}{\|A^{-1}\|_2},$$

where the maximum is attained when z is any nonzero multiple of y, and $A+\delta A$ is singular because

$$(A + \delta A)y = Ay - \frac{xy^T y}{\|A^{-1}\|_2} = \frac{x}{\|A^{-1}\|_2} - \frac{x}{\|A^{-1}\|_2} = 0 \,. \quad \square$$

We have now seen that the distance to the nearest ill-posed problem equals the reciprocal of the condition number for two problems: polynomial evaluation and linear equation solving. This reciprocal relationship is quite common in numerical analysis [71].

Here is a slightly different way to do perturbation theory for $Ax = b$; we will need it to derive practical error bounds later in section 2.4.4. If $\hat{x}$ is any vector, we can bound the difference $\delta x \equiv \hat{x} - x = \hat{x} - A^{-1}b$ as follows. We let $r = A\hat{x} - b$ be the *residual* of $\hat{x}$; the residual r is zero if $\hat{x} = x$. This lets us write $\delta x = A^{-1}r$, yielding the bound

$$\|\delta x\| = \|A^{-1}r\| \leq \|A^{-1}\| \cdot \|r\|. \tag{2.5}$$

This simple bound is attractive to use in practice, since r is easy to compute, given an approximate solution $\hat{x}$. Furthermore, there is no apparent need to estimate δA and δb. In fact our two approaches are very closely related, as shown by the next theorem.

THEOREM 2.2. *Let $r = A\hat{x} - b$. Then there exists a δA such that $\|\delta A\| = \frac{\|r\|}{\|\hat{x}\|}$ and $(A + \delta A)\hat{x} = b$. No δA of smaller norm and satisfying $(A + \delta A)\hat{x} = b$ exists. Thus, δA is the smallest possible backward error (measured in norm). This is true for any vector norm and its induced norm (or $\|\cdot\|_2$ for vectors and $\|\cdot\|_F$ for matrices).*

Proof. $(A+\delta A)\hat{x} = b$ if and only if $\delta A \cdot \hat{x} = b - A\hat{x} = -r$, so $\|r\| = \|\delta A \cdot \hat{x}\| \leq \|\delta A\| \cdot \|\hat{x}\|$, implying $\|\delta A\| \geq \frac{\|r\|}{\|\hat{x}\|}$. We complete the proof only for the two-norm and its induced matrix norm. Choose $\delta A = \frac{-r \cdot \hat{x}^T}{\|\hat{x}\|_2^2}$. We can easily verify that $\delta A \cdot \hat{x} = -r$ and $\|\delta A\|_2 = \frac{\|r\|_2}{\|\hat{x}\|_2}$. □

Thus, the smallest $\|\delta A\|$ that could yield an $\hat{x}$ satisfying $(A+\delta A)\hat{x} = b$ and $r = A\hat{x} - b$ is given by Theorem 2.2. Applying error bound (2.2) (with $\delta b = 0$) yields

$$\|\delta x\| \leq \|A^{-1}\| \left(\frac{\|r\|}{\|\hat{x}\|} \cdot \|\hat{x}\| \right) = \|A^{-1}\| \cdot \|r\|,$$

the same bound as (2.5).

All our bounds depend on the ability to estimate the condition number $\|A\| \cdot \|A^{-1}\|$. We return to this problem in section 2.4.3. Condition number estimates are computed by LAPACK routines such as **sgesvx**.

2.2.1. Relative Perturbation Theory

In the last section we showed how to bound the norm of the error $\delta x = \hat{x} - x$ in the approximate solution $\hat{x}$ of $Ax = b$. Our bound on $\|\delta x\|$ was proportional to the condition number $\kappa(A) = \|A\| \cdot \|A^{-1}\|$ times the norms $\|\delta A\|$ and $\|\delta b\|$, where $\hat{x}$ satisfies $(A + \delta A)\hat{x} = b + \delta b$.

In many cases this bound is quite satisfactory, but not always. Our goal in this section is to show when it is too pessimistic and to derive an alternative perturbation theory that provides tighter bounds. We will use this perturbation theory later in section 2.5.1 to justify the error bounds computed by the LAPACK subroutines like **sgesvx**.

This section may be skipped on a first reading.

Here is an example where the error bound of the last section is much too pessimistic.

EXAMPLE 2.1. Let $A = \text{diag}(\gamma, 1)$ (a diagonal matrix with entries $a_{11} = \gamma$ and $a_{22} = 1$) and $b = [\gamma, 1]^T$, where $\gamma > 1$. Then $x = A^{-1}b = [1, 1]^T$. Any reasonable direct method will solve $Ax = b$ very accurately (using two divisions b_i/a_{ii}) to get $\hat{x}$, yet the condition number $\kappa(A) = \gamma$ may be arbitrarily large. Therefore our error bound (2.3) may be arbitrarily large.

The reason that the condition number $\kappa(A)$ leads us to overestimate the error is that bound (2.2), from which it comes, assumes that δA is bounded in norm *but is otherwise arbitrary*; this is needed to prove that bound (2.2) is attainable in Question 2.3. In contrast, the δA corresponding to the actual rounding errors is not arbitrary but has a special structure not captured by its norm alone. We can determine the smallest δA corresponding to $\hat{x}$ for our problem as follows: A simple rounding error analysis shows that $\hat{x}_i = (b_i/a_{ii})/(1 + \delta_i)$, where $|\delta_i| \leq \varepsilon$. Thus $(a_{ii} + \delta_i a_{ii})\hat{x}_i = b_i$. We may rewrite this

```
issparse(A)
B = full(A);                    % nonsparse version of A
issparse(B)
tic; c1 = condest(A), toc
tic; c2 = cond(B,1), toc
```

Comment on the speed and accuracy of `condest`. You might also like to try `tic; c3 = cond(B,2), toc`, which computes $\kappa_2(B)$. However, you may find that this takes too long unless you decrease the size of the problem by decreasing m. This is a time consuming calculation, because it requires the singular values of B (Section 4.2). □

MATLAB's `condest` function has to compute the LU decomposition of the matrix, since it cannot assume that an LU decomposition is available. Thus `condest` is less efficient than it would be if the decomposition were assumed available. However, if the matrix under consideration is sparse, `condest` will do a sparse LU decomposition (Section 1.9), thereby saving a lot of work. This explains the good outcome in Exercise 2.2.30.

2.3 PERTURBING THE COEFFICIENT MATRIX

Up to this point we have considered only the effect of perturbing b in the system $Ax = b$. We must also consider perturbations of A, as A is also known and represented only approximately. Thus, let us compare two systems $Ax = b$ and $(A + \delta A)\hat{x} = b$, where $\|\delta A\|/\|A\|$ is small. Our first task is to establish a condition that guarantees that the system$(A + \delta A)\hat{x} = b$ has a unique solution, given that the system $Ax = b$ does. This is given by the following theorem, which, along with the subsequent theorems in this section, is valid for any vector norm and its induced matrix norm and condition number.

Theorem 2.3.1 *If A is nonsingular and*

$$\frac{\|\delta A\|}{\|A\|} < \frac{1}{\kappa(A)}, \tag{2.3.2}$$

then $A + \delta A$ is nonsingular.

Proof. The hypothesis $\|\delta A\|/\|A\| < 1/\kappa(A)$ can be rewritten in various ways, for example, $\|\delta A\| < 1/\|A^{-1}\|$ and $\|\delta A\|\,\|A^{-1}\| < 1$. We'll use this last form of the inequality, and we'll prove the contrapositive form of the theorem: If $A + \delta A$ is singular, then $\|\delta A\|\,\|A^{-1}\| \geq 1$.

Suppose $A + \delta A$ is singular. Then, by Theorem 1.2.3, there is a nonzero vector y such that $(A + \delta A)y = 0$. Reorganizing this equation, we obtain $y = -A^{-1}\delta A y$, which implies $\|y\| = \|A^{-1}\delta A y\| \leq \|A^{-1}\|\,\|\delta A\|\,\|y\|$. Since $\|y\| > 0$, we can divide both sides of the inequality by $\|y\|$ to obtain $1 \leq \|A^{-1}\|\,\|\delta A\|$, which is the desired result. □

Theorem 2.3.1 demonstrates another important function of the condition number; it gives us an idea of the distance from A to the nearest nonsingular matrix: If $A+\delta A$ is singular, then $\|\delta A\|/\|A\|$ must be at least $1/\kappa(A)$. It turns out that for the spectral norm this result is exact: If $A+\delta A$ is the singular matrix closest to A, in the sense that $\|\delta A\|_2$ is as small as possible, then $\|\delta A\|_2/\|A\|_2$ is exactly $1/\kappa_2(A)$. We will prove this in Corollary 4.2.22.

As long as (2.3.2) is satisfied, we are assured that the equation $(A+\delta A)\hat{x}=b$ has a unique solution. Notice that (2.3.2) is hard to satisfy if A is ill conditioned; that is, it is satisfied only for very small perturbations δA. If, on the other hand, A is well conditioned, (2.3.2) holds even for relatively large perturbations.

Now let us consider the relationship between the solutions of $Ax=b$ and $(A+\delta A)\hat{x}=b$. Let $\delta x=\hat{x}-x$, so that $\hat{x}=x+\delta x$. Under what conditions can we conclude that $\|\delta x\|/\|x\|$ is small? We would like an upper bound on $\|\delta x\|/\|x\|$ in the spirit of Theorem 2.2.4. We will obtain such a bound eventually, but it turns out to be easier to bound $\|\delta x\|/\|\hat{x}\|$. In most cases there will not be much difference between $\|x\|$ and $\|\hat{x}\|$, so it makes little difference which one we use in the denominator.

Theorem 2.3.3 *Let A be nonsingular, let $b\neq 0$, and let x and $\hat{x}=x+\delta x$ be solutions of $Ax=b$ and $(A+\delta A)\hat{x}=b$, respectively. Then,*

$$\frac{\|\delta x\|}{\|\hat{x}\|}\leq \kappa(A)\frac{\|\delta A\|}{\|A\|}. \tag{2.3.4}$$

Proof. Rewriting the equation $(A+\delta A)\hat{x}=b$ as $Ax+A\delta x+\delta A\hat{x}=b$, using the equation $Ax=b$, and reorganizing the resulting equation, we obtain $\delta x=-A^{-1}\delta A\hat{x}$. Thus

$$\|\delta x\|\leq \|A^{-1}\|\,\|\delta A\|\,\|\hat{x}\|. \tag{2.3.5}$$

Dividing through by $\|\hat{x}\|$ and using the definition $\kappa(A)=\|A\|\,\|A^{-1}\|$, we obtain the desired result. □

Theorem 2.3.3 shows that once again the condition number of A plays the decisive role. If $\kappa(A)$ is not too large, then a small perturbation in A results in a small perturbation in x, in the sense that $\|\delta x\|/\|\hat{x}\|$ is small.

It is interesting to note that Theorem 2.3.3 does not rely on nonsingularity of $A+\delta A$, nor on any assumption to the effect that δA is small. In contrast, the next theorem, which provides a bound on $\|\delta x\|/\|x\|$, does make such an assumption.

Theorem 2.3.6 *If A is nonsingular, $\|\delta A\|/\|A\|<1/\kappa(A)$, $b\neq 0$, $Ax=b$, and $(A+\delta A)(x+\delta x)=b$, then*

$$\frac{\|\delta x\|}{\|x\|}\leq \frac{\kappa(A)\dfrac{\|\delta A\|}{\|A\|}}{1-\kappa(A)\dfrac{\|\delta A\|}{\|A\|}}. \tag{2.3.7}$$

If A is well conditioned and $\|\delta A\|/\|A\|$ is sufficiently small, then $\|\delta A\|/\|A\| \ll 1/\kappa(A)$. In this case the denominator on the right side of (2.3.7) is approximately 1. Then (2.3.7) states *roughly* that

$$\frac{\|\delta x\|}{\|x\|} \leq \kappa(A)\frac{\|\delta A\|}{\|A\|},$$

which is almost the same as (2.3.4). This shows that if A is well conditioned and $\|\delta A\|/\|A\|$ is small, then $\|\delta x\|/\|x\|$ is small.

If, on the other hand, A is ill conditioned, then (2.3.7) allows (but does not prove) that $\|\delta x\|/\|x\|$ could be large, even if $\|\delta A\|/\|A\|$ is small.

Proof. The proof of Theorem 2.3.6 is the same as that of Theorem 2.3.3, up to (2.3.5). Rewriting $\hat{x}$ as $x + \delta x$ in (2.3.5) and using the triangle inequality, we find that

$$\begin{aligned}\|\delta x\| &\leq \|A^{-1}\|\,\|\delta A\|\,(\|x\| + \|\delta x\|) \\ &= \kappa(A)\frac{\|\delta A\|}{\|A\|}\,(\|x\| + \|\delta x\|)\,.\end{aligned}$$

Now rewrite this inequality so that all of the terms involving $\|\delta x\|$ are on the left-hand side.

$$\left(1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}\right)\|\delta x\| \leq \kappa(A)\frac{\|\delta A\|}{\|A\|}\|x\|$$

The assumption $\|\delta A\|/\|A\| < 1/\kappa(A)$ guarantees that the factor that multiplies $\|\delta x\|$ is positive, so we can divide by it without reversing the inequality. If we also divide through by $\|x\|$, we obtain the desired result. □

So far we have considered the effects of perturbing b and A separately. This was done not out of necessity but from a desire to keep the analysis simple. The combined effects of perturbations in A and b can be expressed in a single inequality, as the next two theorems show. The first is in the spirit of Theorem 2.3.3, and the second is in that of Theorem 2.3.6.

Theorem 2.3.8 *Let A be nonsingular, and suppose x and $\hat{x}$ satisfy $Ax = b$ and $\hat{A}\hat{x} = \hat{b}$, respectively, where $\hat{A} = A + \delta A$, $\hat{x} = x + \delta x \neq 0$, and $\hat{b} = b + \delta b \neq 0$. Then*

$$\frac{\|\delta x\|}{\|\hat{x}\|} \leq \kappa(A)\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|\hat{b}\|} + \frac{\|\delta A\|}{\|A\|}\frac{\|\delta b\|}{\|\hat{b}\|}\right).$$

Of the terms on the right-hand side, the product term is usually negligible. For example, if $\frac{\|\delta A\|}{\|A\|} = 10^{-5}$ and $\frac{\|\delta b\|}{\|\hat{b}\|} = 10^{-5}$, then $\frac{\|\delta A\|}{\|A\|}\frac{\|\delta b\|}{\|\hat{b}\|} = 10^{-10}$.

Theorem 2.3.9 *If A is nonsingular, $\|\delta A\|/\|A\| < 1/\kappa(A)$, $b \neq 0$, $Ax = b$, and $(A + \delta A)(x + \delta x) = b + \delta b$, then*

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa(A)\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)}{1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}}. \tag{2.3.10}$$

Example 2.3.11 In Example 1.2.6 we considered an electrical circuit that leads to the linear system

$$\begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 1.5 & 0 & -.5 \\ -1 & 0 & 1.7 & -.2 \\ 0 & -.5 & -.2 & 1.7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 3 \\ 0 \end{bmatrix},$$

which we solved to determine the voltages at the nodes of the circuit. If we solve the system using MATLAB or any other reliable software, we obtain an extremely accurate solution (See Example 2.4.2 below). It is the accurate solution of the given system, but what if the entries of A and b are incorrect? The entries of A depend on the resistances in the circuit, and the one nonzero entry of b depends also on the voltage of the battery. None of these quantities are known exactly. Theorem 2.3.9 gives us information about the effects of inaccuracies. Suppose, for example, the resistances and the voltage are known to be in error by less than one one hundredth of one percent. This means that the relative error is less than 10^{-4}, so, roughly speaking,

$$\frac{\|\delta A\|_2}{\|A\|_2} \approx 10^{-4} \quad \text{and} \quad \frac{\|\delta b\|_2}{\|b\|_2} \approx 10^{-4}.$$

Using MATLAB's `cond` function, we get $\kappa_2(A) = 12.7$. Substituting these values into (2.3.10), we find that

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq 2.5 \times 10^{-3}.$$

Thus the computed nodal voltages are off by at most one quarter of one percent. It should be noted that the actual error is likely to be much less than this. Results obtained using an upper bound like the one in Theorem 2.3.9 tend to be quite pessimistic. □

Exercise 2.3.12 Prove Theorem 2.3.8. Do it your way, or use the following outline.

(a) Show that

$$\delta x = A^{-1}(\delta b - \delta A \hat{x}),$$

$$\|\delta x\| \leq \|A^{-1}\| (\|\delta b\| + \|\delta A\| \, \|\hat{x}\|),$$

and

$$\frac{\|\delta x\|}{\|\hat{x}\|} \leq \kappa(A) \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \, \|\hat{x}\|} \right).$$

(b) Show that

$$\|\hat{b}\| \leq (\|A\| + \|\delta A\|) \, \|\hat{x}\|,$$

and therefore

$$\frac{1}{\|\hat{x}\|} \leq \frac{(\|A\| + \|\delta A\|)}{\|\hat{b}\|}.$$

(c) Combine the results of (a) and (b) to finish the proof.