ESTIMATION OF ERROR

Let $\widehat{x}$ denote an approximate solution for $Ax = b$; perhaps $\widehat{x}$ is obtained by Gaussian elimination. Let x denote the exact solution. Then introduce

$$r = b - A\widehat{x}$$

a quantity called the *residual* for $\widehat{x}$. Then

$$\begin{aligned} r &= b - A\widehat{x} \\ &= Ax - A\widehat{x} \\ &= A(x - \widehat{x}) \\ x - \widehat{x} &= A^{-1}r \end{aligned}$$

or the error $e = x - \widehat{x}$ is the exact solution of

$$Ae = r$$

Thus we can solve this to obtain an estimate $\widehat{e}$ of our error e.

EXAMPLE. Recall the linear system

$$\begin{array}{rcrcrcl} .729x_1 & + & .81x_2 & + & .9x_3 & = & .6867 \\ x_1 & + & x_2 & + & x_3 & = & .8338 \\ 1.331x_1 & + & 1.21x_2 & + & 1.1x_3 & = & 1.000 \end{array}$$

The true solution, rounded to four significant digits, is

$$x = [.2245, .2814, .3279]^{\mathsf{T}}$$

Using Gaussian elimination without pivoting and four digit decimal floating point arithmetic with rounding, the resulting solution and error are

$$\widehat{x} = [.2251, .2790, .3295]^{\mathsf{T}}$$
$$e = [-.0006, .0024, -.0016]^{\mathsf{T}}$$

Then

$$r = b - A\widehat{x} = [.00006210, .0002000, .0003519]^{\mathsf{T}}$$

Solving $Ae = r$ by Gaussian elimination, we obtain

$$e \approx \widehat{e} = [-.0004471, .002150, -.001504]^{\mathsf{T}}$$

THE RESIDUAL CORRECTION METHOD

If in the above we had taken $\widehat{e}$ and added it to $\widehat{x}$, then we would have obtained an improved answer:

$$x \approx \widehat{x} + \widehat{e} = [.2247, .2811, .3280]^{\mathsf{T}}$$

Recall

$$x = [.2245, .2814, .3279]^{\mathsf{T}}$$

With the new approximation, we can repeat the earlier process of estimating the error and then using it to improve the answer. This iterative process is called the *residual correction method*. It is illustrated with another example on page 297 in the text.

ERROR ANALYSIS

Begin with a simple example. The system

$$\begin{array}{rcrcl} 7x & + & 10y & = & 1 \\ 5x & + & 7y & = & .7 \end{array}$$

has the solution

$$x = 0, \quad y = .1$$

The perturbed system

$$\begin{array}{rcrcr} 7\widehat{x} & + & 10\widehat{y} & = & 1.01 \\ 5\widehat{x} & + & 7\widehat{y} & = & .69 \end{array}$$

has the solution

$$\widehat{x} = -.17, \quad \widehat{y} = .22$$

Why is there such a difference?

Consider the following Hilbert matrix example.

$$H_3 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}, \quad \widetilde{H}_3 = \begin{bmatrix} 1.000 & .5000 & .3333 \\ .5000 & .3333 & .2500 \\ .3333 & .2500 & .2000 \end{bmatrix}$$

$$H_3^{-1} = \begin{bmatrix} 9 & -36 & 30 \\ -36 & 192 & -180 \\ 30 & -180 & 180 \end{bmatrix}$$

$$\widetilde{H}_3^{-1} = \begin{bmatrix} 9.062 & -36.32 & 30.30 \\ -36.32 & 193.7 & -181.6 \\ 30.30 & -181.6 & 181.5 \end{bmatrix}$$

We have changed H_3 in the fifth decimal place (by rounding the fractions to four decimal digits). But we have ended with a change in H_3^{-1} in the third decimal place.

VECTOR NORMS

A *norm* is a generalization of the absolute value function, and we use it to measure the “size” of a vector. There are a variety of ways of defining the norm of a vector, with each definition tied to certain applications.

Euclidean norm : Let x be a column vector with n components. Define

$$\|x\|_2 = \left[\sum_{j=1}^{n} \left|x_j\right|^2\right]^{\frac{1}{2}}$$

This is the standard definition of the length of a vector, giving us the “straight line” distance between head and tail of the vector.

1-norm : Let x be a column vector with n components. Define

$$\|x\|_1 = \sum_{j=1}^{n} \left|x_j\right|$$

For planar applications ($n = 2$), this is sometimes called the "taxi cab norm", as it corresponds to distance as measured when driving in a city laid out with a rectangular grid of streets.

∞*-norm* : Let x be a column vector with n components. Define

$$\|x\|_\infty = \max_{1 \le j \le n} \left|x_j\right|$$

This is also called the *maximum norm* and the *Chebyshev norm*. It is often used in numerical analysis where we want to measure the maximum error component in some vector quantity.

EXAMPLES

Let

$$x = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}^{\mathsf{T}}$$

Then

$$\begin{aligned} \|x\|_1 &= 6 \\ \|x\|_2 &= \text{sqrt}(14) \doteq 3.74 \\ \|x\|_\infty &= 3 \end{aligned}$$

PROPERTIES

Let $\|\cdot\|$ denote a generic norm. Then:

(a) $\|x\| = 0$ if and only if $x = 0$.
(b) $\|cx\| = |c| \, \|x\|$ for any vector x and constant c.
(c) $\|x + y\| \leq \|x\| + \|y\|$, for all vectors x and y.

MATRIX NORMS

We also need to measure the sizes of general matrices, and we need to have some way of relating the sizes of A and x to the size of Ax. In doing this, we will consider only square matrices A.

We say a matrix norm is a way of defining the size of a matrix, again satisfying the properties seen with vector norms. Thus:

1. $\|A\| = 0$ if and only if $A = 0$.

2. $\|cA\| = |c| \, \|A\|$ for any matrix A and constant c.

3. $\|A + B\| \le \|A\| + \|B\|$, for all matrices A and B of equal order.

In addition, we can multiply matrices, forming AB from A and B. With absolute values, we have $|ab| = |a|\,|b|$ for all complex numbers a and b. There is no way of generalizing exactly this relation to matrices A and B. But we can obtain definitions for which

$$\text{(d)} \qquad \|AB\| \leq \|A\|\,\|B\|$$

Finally, if we are given some vector norm $\|\cdot\|_v$, we can obtain an associated matrix norm definition for which

$$\text{(e)} \qquad \|Ax\|_v \leq \|A\|\,\|x\|_v$$

for all $n \times n$ matrices A and $n \times 1$ vectors x.

Often we use as our definition of $\|A\|$ the smallest number for which this last inequality is satisfied for all vectors x. In that case, we also obtain the useful property

$$\|I\| = 1$$

Let the vector norm be $\|\cdot\|_\infty$ for $n\times 1$ vectors x. Then the associated matrix norm definition is

$$\|A\| = \max_{1\le i\le n} \sum_{j=1}^{n} \left|a_{i,j}\right|$$

This is sometimes called the "row norm" of a matrix A.

EXAMPLE. Let

$$A = \begin{bmatrix} 1 & 2 \\ 5 & 7 \end{bmatrix}, \quad z = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad Az = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

Then

$$\|A\| = 12, \quad \|z\|_\infty = 1, \quad \|Az\|_\infty = 2$$

and clearly $\|Az\|_\infty \le \|A\|\,\|z\|_\infty$. Also let $z = \begin{bmatrix} 1, & 1 \end{bmatrix}^{\mathsf{T}}$. Then

$$Az = \begin{bmatrix} 3 \\ 12 \end{bmatrix}, \quad \|z\|_\infty = 1, \quad \|Az\|_\infty = 12$$

and $\|Az\|_\infty = \|A\|\,\|z\|_\infty$.

Let the vector norm be $\|\cdot\|_1$. Then the associated matrix norm definition is

$$\|A\| = \max_{1 \le j \le n} \sum_{i=1}^{n} \left| a_{i,j} \right|$$

This is sometimes called the "column norm" of the matrix A.

Let the vector norm be $\|\cdot\|_2$. Then the associated matrix norm definition is

$$\|A\| = \text{sqrt}\left[r_\sigma(A^{\mathsf{T}} A) \right]$$

To understand this, let B denote an arbitrary square matrix of order $n \times n$. Then introduce

$$\sigma(B) = \{\lambda \text{ an eigenvalue of } B\}$$

$$r_\sigma(B) = \max_{\lambda \in \sigma(B)} |\lambda|$$

The set $\sigma(B)$ is called the spectrum of B, and it contains all the eigenvalues of B. The number $r_\sigma(B)$ is called the "spectral radius" of B. There are easily computable bounds for $\|A\|$, but the norm itself is difficult to compute.

ERROR BOUNDS

Let $Ax = b$ and $A\widehat{x} = \widehat{b}$, and we are interested in cases with $b \approx \widehat{b}$. Then

$$\frac{\|x - \widehat{x}\|_v}{\|x\|_v} \leq \|A\| \left\|A^{-1}\right\| \frac{\left\|b - \widehat{b}\right\|_v}{\|b\|_v}$$

where $\|\cdot\|_v$ is some vector norm and $\|\cdot\|$ is an associated matrix norm.

Proof:

$$\begin{aligned}
Ax - A\widehat{x} &= b - \widehat{b} \\
A(x - \widehat{x}) &= b - \widehat{b} \\
x - \widehat{x} &= A^{-1}\left(b - \widehat{b}\right) \\
\|x - \widehat{x}\|_v &= \left\|A^{-1}\left(b - \widehat{b}\right)\right\|_v \\
&\leq \left\|A^{-1}\right\| \left\|b - \widehat{b}\right\|_v \\
\frac{\|x - \widehat{x}\|_v}{\|x\|_v} &\leq \frac{\left\|A^{-1}\right\| \left\|b - \widehat{b}\right\|_v}{\|x\|_v}
\end{aligned}$$

Rewrite this as

$$\frac{\|x - \widehat{x}\|_v}{\|x\|_v} \le \|A\| \left\|A^{-1}\right\| \frac{\left\|b - \widehat{b}\right\|_v}{\|A\| \, \|x\|_v}$$

Since $Ax = b$, we have

$$\|b\|_v = \|Ax\|_v \le \|A\| \, \|x\|_v$$

Using this,

$$\frac{\|x - \widehat{x}\|_v}{\|x\|_v} \le \|A\| \left\|A^{-1}\right\| \frac{\left\|b - \widehat{b}\right\|_v}{\|b\|_v}$$

This completes the proof of the earlier assertion.

The quantity

$$\text{cond}(A) = \|A\| \left\|A^{-1}\right\|$$

is called a <u>condition number</u> for the matrix A.

EXAMPLE. Recall the earlier example:

$$\begin{array}{rcrcr} 7x_1 & + & 10x_2 & = & 1 \\ 5x_1 & + & 7x_2 & = & .7 \end{array} \qquad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ .1 \end{bmatrix}$$

$$\begin{array}{rcrcr} 7\widehat{x}_1 & + & 10\widehat{x}_2 & = & 1.01 \\ 5\widehat{x}_1 & + & 7\widehat{x}_2 & = & .69 \end{array} \qquad \begin{bmatrix} \widehat{x}_1 \\ \widehat{x}_2 \end{bmatrix} = \begin{bmatrix} -.17 \\ .22 \end{bmatrix}$$

Then

$$\|b\|_\infty = 1, \quad \left\|b - \widehat{b}\right\|_\infty = .01$$

$$\|x\|_\infty = .1, \quad \|x - \widehat{x}\|_\infty = .17$$

$$A = \begin{bmatrix} 7 & 10 \\ 5 & 7 \end{bmatrix}, \qquad A^{-1} = \begin{bmatrix} -7 & 10 \\ 5 & -7 \end{bmatrix}$$

$$\|A\| = 17, \qquad \left\|A^{-1}\right\| = 17, \qquad \text{cond}(A) = 289$$

$$\frac{\|x - \widehat{x}\|_\infty}{\|x\|_\infty} \div \frac{\left\|b - \widehat{b}\right\|_\infty}{\|b\|_\infty} = \frac{1.7}{.01} = 170 \le \text{cond}(A)$$

$$\frac{\|x - \widehat{x}\|_\infty}{\|x\|_\infty} \le \text{cond}(A) \frac{\left\|b - \widehat{b}\right\|_\infty}{\|b\|_\infty}$$

The result

$$\frac{\|x - \widehat{x}\|_v}{\|x\|_v} \le \text{cond}(A) \frac{\left\|b - \widehat{b}\right\|_v}{\|b\|_v}$$

has another aspect which we do not prove here. Given any matrix A, then there is a vector b and a nearby perturbation $\widehat{b}$ for which the above inequality can be replaced by equality. Moreover, there is no simple way to know of these b and $\widehat{b}$ in advance. For such b and $\widehat{b}$, we have

$$\text{cond}(A) = \frac{\|x - \widehat{x}\|_v}{\|x\|_v} \div \frac{\left\|b - \widehat{b}\right\|_v}{\|b\|_v}$$

Thus if cond(A) is very large, say 10^8, then there are b and $\widehat{b}$ for which

$$\frac{\|x - \widehat{x}\|_v}{\|x\|_v} = 10^8 \cdot \frac{\left\|b - \widehat{b}\right\|_v}{\|b\|_v}$$

We call such systems *ill-conditioned*.

Recall an earlier discussion of error in Gaussian elimination. Let $\widehat{x}$ denote an approximate solution for $Ax = b$; perhaps $\widehat{x}$ is obtained by Gaussian elimination. Let x denote the exact solution. Then introduce the *residual*

$$r = b - A\widehat{x}$$

We then obtained $x - \widehat{x} = A^{-1}r$. But we could also have discussed this as a special case of our present results. Write

$$Ax = b \quad \text{and} \quad A\widehat{x} = b - r \equiv \widehat{b}$$

Then

$$\frac{\|x - \widehat{x}\|_v}{\|x\|_v} \leq \text{cond}(A) \frac{\left\|b - \widehat{b}\right\|_v}{\|b\|_v}$$

becomes

$$\frac{\|x - \widehat{x}\|_v}{\|x\|_v} \leq \text{cond}(A) \frac{\|r\|_v}{\|b\|_v}$$

ILL-CONDITIONED EXAMPLE

Define the 4×4 Hilbert matrix:

$$H_4 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{bmatrix}$$

Its inverse is given by

$$H_4^{-1} = \begin{bmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{bmatrix}$$

For the matrix row norm,

$$\text{cond}(H_4) = \frac{25}{12} \cdot 13620 = 28375$$

Thus rounding error in defining b should lead to errors in solving $H_4 x = b$ that are larger than the rounding errors by a factor of 10^4 or more.