ERROR ANALYSIS

Begin with a simple example. The system

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

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.

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

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

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

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.

In solving a linear system $Ax = b$, we need to know the sensitivity of the solution x to changes in the right side b. Consider the linear systems

$$Ax = b, \quad A\widetilde{x} = b + r$$

What is

$$\frac{\|\widetilde{x} - x\|}{\|x\|}\text{?}$$

We simply solve for $\widetilde{x} - x$:

$$\begin{aligned} \widetilde{x} - x &= A^{-1}[b + r] - A^{-1}b \\ &= A^{-1}r \\ \|\widetilde{x} - x\| &= \left\|A^{-1}r\right\| \\ &\leq \left\|A^{-1}\right\| \|r\| \end{aligned}$$

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \leq \frac{\left\|A^{-1}\right\| \|r\|}{\|x\|} = \left\|A^{-1}\right\| \|A\| \frac{\|r\|}{\|A\| \|x\|}$$

Since $Ax = b$, we have $\|b\| \leq \|A\| \|x\|$, and then

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \leq \left\|A^{-1}\right\| \|A\| \frac{\|r\|}{\|b\|}$$

The number

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

is called a *condition number* for the matrix A and the linear system $Ax = b$.

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

We can also prove a lower inequality, obtaining

$$\frac{1}{\text{cond}(A)}\frac{\|r\|}{\|b\|} \leq \frac{\|\widetilde{x} - x\|}{\|x\|} \leq \text{cond}(A)\frac{\|r\|}{\|b\|}$$

In addition, given any nonsingular A, there are vectors b and r for which either of the above inequalities is actually an equality.

With this,

$$\frac{1}{\mathrm{cond}(A)}\frac{\|r\|}{\|b\|} \leq \frac{\|\widetilde{x} - x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|r\|}{\|b\|}$$

we see that if $\mathrm{cond}(A) \approx 1$, then small 'relative' changes in b are guaranteed to lead to equally small 'relative' changes in x.

But if $\mathrm{cond}(A)$ is very large, then there are values of b and r for which

$$\frac{\|r\|}{\|b\|}$$

is small and

$$\frac{\|\widetilde{x} - x\|}{\|x\|}$$

is large.In practice, it is very difficult to know whether your choice of b and r is good or bad.

A LOWER BOUND

Recall that for any matrix norm,

$$r_\sigma(A) \le \|A\|\,, \quad r_\sigma(A^{-1}) \le \left\|A^{-1}\right\|$$

Thus

$$r_\sigma(A) r_\sigma(A^{-1}) \le \text{cond}(A)$$

Also, the eigenvalues of A^{-1} are the reciprocals of those of A. Thus

$$\text{cond}(A)_* \equiv \frac{\max\limits_{\lambda\in\sigma(A)} |\lambda|}{\min\limits_{\lambda\in\sigma(A)} |\lambda|} \le \text{cond}(A)$$

This ratio of eigenvalues is also often used as a condition number, and it is further illustrated in the text (on page 532).

GASTINEL'S THEOREM

Let A be any nonsingular matrix, and let $\|\cdot\|$ denote an operator matrix norm. Then

$$\frac{1}{\text{cond}(A)} = \min\left\{ \frac{\|A - B\|}{\|A\|} \middle| \, B \text{ is a singular matrix} \right\}$$

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

This is equivalent to saying

$$\frac{1}{\left\|A^{-1}\right\|} = \min\{\|P\| : A + P \text{ is singular}\}$$

The first result says that the reciprocal of cond(A) is a measure of how close A is to singular matrix in a "relative error sense".

A PERTURBATION THEOREM

Let $Ax = b$ denote a nonsingular linear system. We perturb both the right side b and the matrix A, asking the consequence on the accuracy of the solution x. In particular, consider the perturbed system

$$(A + \delta A)\,\widetilde{x} = b + \delta b$$

with δA and δb denoting a matrix and vector of small size. What is $x - \widetilde{x}$? We will introduce $\widetilde{x} = x + \delta x$, thus having

$$(A + \delta A)\,(x + \delta x) = b + \delta b$$

First: Is $A + \delta A$ nonsingular? We write it as

$$A + \delta A = A\left[I + A^{-1}\delta A\right]$$

If we assume

$$\|\delta A\| < \frac{1}{\left\|A^{-1}\right\|} \tag{1}$$

then $\left[I + A^{-1}\delta A\right]^{-1}$ exists by the geometric series theorem, with

$$\left\|\left[I + A^{-1}\delta A\right]^{-1}\right\| \le \frac{1}{1 - \|\delta A\| \left\|A^{-1}\right\|}$$

Note the condition (1) is equivalent to writing

$$\frac{\|\delta A\|}{\|A\|} < \frac{1}{\text{cond}(A)}$$

Thus for a matrix with a large condition number, only very small perturbations of A are allowed in our analysis.

Assuming (1), we have that the perturbed equation

$$(A + \delta A)(x + \delta x) = b + \delta b$$

is uniquely solvable. Subtract from it $Ax = b$ and then do some algebraic manipulations, to obtain

$$\left[I + A^{-1}(\delta A)\right]\delta x = A^{-1}(\delta b) - A^{-1}(\delta A)\, x$$

$$\left[I + A^{-1}(\delta A)\right]\delta x = A^{-1}(\delta b) - A^{-1}(\delta A)\,x$$

$$\delta x = \left[I + A^{-1}(\delta A)\right]^{-1}\left[A^{-1}(\delta b) - A^{-1}(\delta A)\,x\right]$$

$$\|\delta x\| \leq \left\|\left[I + A^{-1}(\delta A)\right]^{-1}\right\| \left\|A^{-1}\right\| [\|\delta b\| + \|\delta A\|\,\|x\|]$$

Following further manipulations, given in the book, we have

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \text{cond}(A)\frac{\|\delta A\|}{\|A\|}}\left[\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right] \qquad (2)$$

This shows that if cond(A) is relatively small, say $1 \leq \text{cond}(A) \leq 10$, then small relative changes in the data (A and b) will lead to small relative changes in the solution. And as before, if cond(A) is large, then there are choices of b and δb for which the relative change in x will be very large.

A BEST POSSIBLE RESULT

Assume the entire Gaussian Elimination process is carried thru with no error, with the only errors being the initial storage of b and A into the computer memory, where there is expected to be a rounding error. Thus assume

$$fl(b) = b + \delta b, \qquad fl(A) = A + \delta A$$

where

$$\|\delta b\|_\infty \le \mathbf{u} \, \|b\|_\infty \,, \qquad \|\delta A\|_\infty \le \mathbf{u} \, \|A\|_\infty$$

with $\mathbf{u}$ the unit round of the machine. Further assume that

$$(A + \delta A) \, \widehat{x} = b + \delta b$$

and that $\mathbf{u} \, \mathrm{cond}(A) \le \frac{1}{2}$. Then (2) implies

$$\frac{\|x - \widehat{x}\|_\infty}{\|x\|_\infty} \le 4\mathbf{u} \, \mathrm{cond}_\infty(A)$$

$$\frac{\|x - \widehat{x}\|_\infty}{\|x\|_\infty} \le 4\mathbf{u}\,\mathsf{cond}_\infty(A)$$

With every b and A, this is an attainable bound for suitably chosen δb and δA. Every linear system $Ax = b$ will almost certainly have its solution affected to at least this extant, and thus ill-conditioned problems will still be affected a great deal, even with only a minimal occurrence of rounding error.

This idea and discussion is taken from the book
Golub & Van Loan, *Matrix Computations*, 3rd edition,
page 104.

HILBERT MATRIX

The Hilbert matrix is defined by

$$H_n = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \ddots & & \\ \vdots & & & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & & \cdots & \frac{1}{2n-1} \end{bmatrix}$$

It occurs naturally as the coefficient matrix in the computation of the least squares polynomial fit to a given function $f(x)$ on the interval $[0, 1]$ (cf. §4.3, page 207, in the text). It is a very ill-conditioned matrix for larger values of n.

n	cond$_2(A)$
3	$5.24E + 2$
6	$1.50E + 7$
9	$4.93E + 11$

FORWARD ERROR ANALYSIS

A standard way to understand the effects of rounding errors in a calculation is to carry out a *forward error analysis* of the computation. For example, return to the analysis of summation in §1.5 of Chapter 1. We considered the computation of

$$S = \sum_{j=1}^{m} a_j$$

with x_j machine floating point numbers. Then we defined

$$\begin{array}{lllll}
S_2 & = & fl\,(a_1 + a_2) & = & (1 + \varepsilon_2)\,(a_1 + a_2) \\
S_3 & = & fl(S_2 + a_3) & = & (1 + \varepsilon_3)\,(S_2 + a_3) \\
S_4 & = & fl(S_3 + a_4) & = & (1 + \varepsilon_4)\,(S_3 + a_4) \\
 & \vdots & & & \\
S_m & = & fl(S_{m-1} + a_m) & = & (1 + \varepsilon_m)\,(S_{m-1} + a_m)
\end{array}$$

We showed that S_m satisfied

$$\begin{aligned} S - S_m \doteq \; & -a_1\left(\varepsilon_2 + \cdots + \varepsilon_m\right) \\ & -a_2\left(\varepsilon_2 + \cdots + \varepsilon_m\right) \\ & -a_3\left(\varepsilon_3 + \cdots + \varepsilon_m\right) \\ & -a_3\left(\varepsilon_4 + \cdots + \varepsilon_m\right) \\ & \quad\vdots \\ & -a_m\varepsilon_m \end{aligned} \tag{3}$$

The numbers ε_j come from the assumption that floating point addition satisfies

$$fl(a+b) = (1+\varepsilon)\,(a+b)$$

for some ε satisfying

$$\begin{array}{llll} -2^{-N} & \leq \varepsilon & \leq 2^{-N}, & \text{rounded arithmetic} \\ -2^{-N+1} & \leq \varepsilon & \leq 0, & \text{chopped arithmetic} \end{array}$$

In this we assume the arithmetic is binary floating point with N binary digits in the mantissa.

With the above formula (3) for $S - S_m$, we can judge the effects of rounding errors occurring in the summation process. This is an example of forward *error analysis*.

BACKWARD ERROR ANALYSIS

With many problems, we obtain a better idea of the effects of rounding errors by using another approach. As an example, consider solving for some root α of an equation

$$x^q + a_{q-1}x^{q-1} + \cdots + a_1 x + a_0 = 0 \qquad (4)$$

Let $\widehat{\alpha}$ denote an approximate root that we obtain by some means, and assume the error in $\widehat{\alpha}$ is primarily due to rounding. Then it is often possible to show that $\widehat{\alpha}$ is the exact root of a related nearby polynomial equation

$$x^q + \widehat{a}_{q-1}x^{q-1} + \cdots + \widehat{a}_1 x + \widehat{a}_0 = 0 \qquad (5)$$

with known bounds on $a_j - \widehat{a}_j$, $j = 0, ..., q-1$. These bounds are often quite small, around the size of the machine ‘unit round’.

One can then look at the effect on the roots of (4) of such perturbations in the coefficients. Such perturbation analyses were given in §2.9 of the text (cf. (2.9.19)). With some equations (4), the roots of (5) will be close to those of (4) when the perturbations $a_j - \widehat{a}_j$ are small. But with other equations, small perturbations in $a_j - \widehat{a}_j$ can lead to much larger changes in the roots.

GAUSSIAN ELIMINATION
ERROR ANALYSIS

At every step of Gaussian elimination, we have rounding errors. Initially, researchers attempted to do a forward error analysis by going from each step to the next, noting what might have been if no rounding had occurred. In most cases, this led nowhere.

In the late 1950s and early 1960s, James Wilkinson took a new approach to the problem. He showed it was better to work backwards, developing a *backward error analysis*. In particular, for Gaussian elimination with partial pivoting, he showed that the computed solution $\widehat{x}$ to $Ax = b$ is the exact solution to a perturbed system

$$(A + \delta A)\,\widehat{x} = b$$

with

$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} \leq \left\{1.01\left(n^3 + 3n^2\right)\rho u\right\}$$

$$\rho = \frac{1}{\|A\|_\infty} \max_{1 \le i,j,k \le n} \left| a_{i,j}^{(k)} \right|$$

and u the unit round of the computer arithmetic being used.

The preceding perturbation theorem (2) can then be used to bound the error in $\widehat{x}$:

$$\frac{\|\widehat{x} - x\|}{\|x\|} \le \frac{\text{cond}(A)}{1 - \text{cond}(A)\frac{\|\delta A\|}{\|A\|}} \cdot \frac{\|\delta A\|}{\|A\|}$$

We still need to know the size of ρ, but it is computable. With only partial pivoting, ρ can be quite large in theory, to increase exponentially with n.. But in practice, it rarely gets very large. This result shows that for matrices A with a condition number that is not too large, the use of Gaussian elimination is generally safe.

In addition, Wilkinson says that in practice, a better empirical bound is

$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} \le nu$$