Condition Number

Wednesday, October 12, 2005
11:03 AM

from last time, we have the norm relation

$$\|(I-A)^{-1}\| \leq \frac{1}{1-\|A\|} \quad \text{when} \quad \|A\| < 1$$

Suppose we are solving $A\vec{x} = \vec{b}$, and we have an approximation ξ

Now, $A\xi = b$ iff ξ is a solution to the system, but in actuality $A\xi = r$, a residual vector

$$A\xi - Ax = r$$

$$A(\xi - x) = r \quad \Rightarrow \quad \|A\|\,\|x-\xi\| \leq \|r\|$$

$$(\xi - x)\, A^{-1} r$$

$$\frac{\|\vec{r}\|}{\|A\|} \leq \|\xi - x\| \leq \|A^{-1}\|\,\|r\|$$

If $\frac{\|\vec{r}\|}{\|A\|}$ is large, then ξ is a poor approximation; if it's large, then we may or may not have a good approximation because $\|A^{-1}\|$ can be large

$$\frac{\|\xi - x\|}{\|x\|} \qquad A\vec{x} = \vec{b}$$

$$\|A\| \cdot \|x\| \geq \|b\|$$

$$\frac{\|A\|}{\|b\|} \geq \frac{1}{\|x\|}$$

$$\frac{\|\xi - x\|}{\|x\|} \leq \frac{\|A\|}{\|b\|} \cdot \|A^{-1}\| \|r\|$$

The number $\|A\| \cdot \|A^{-1}\| = K(A)$ is called the "condition number" of A

Often we think of a matrix *A* as depending on a parameter:

$$A(\varepsilon)\,\vec{x}(\varepsilon) = \vec{b}$$

The solution vector $x(\varepsilon) = A^{-1}(\varepsilon)\,\vec{b}$

we want to see how A changes as ε changes; a Taylor series looks like

$$= \underbrace{A^{-1}(0)\,b}_{\vec{x}(0)} + \varepsilon \frac{dA^{-1}}{d\varepsilon}\Big]_{\varepsilon=0} \vec{b} \quad + O(\varepsilon^2)$$

$A(\varepsilon)\cdot A^{-1}(\varepsilon) = I$ certainly

$$A(\varepsilon)\cdot\frac{dA^{-1}(\varepsilon)}{d\varepsilon} + \frac{dA(\varepsilon)}{d\varepsilon}\cdot A^{-1}(\varepsilon) = 0$$

By chain rule

$$\frac{dA^{-1}(\varepsilon)}{d\varepsilon} = -A^{-1}(\varepsilon)\frac{dA(\varepsilon)}{d\varepsilon}A^{-1}(\varepsilon)$$

Now, say $A(\varepsilon) = A + \varepsilon E, \quad \|E\| < 1$

$dA(\varepsilon)/d\varepsilon = E$

$$\vec{x}(\varepsilon) = \vec{x}(0) + \varepsilon\left(-A^{-1}(0) E A^{-1}(0)\right)\vec{b} + O(\varepsilon^2)$$

$$\|\vec{x}(\varepsilon) - \vec{x}(0)\| \leq |\varepsilon| \cdot \|A^{-1}\| \, \|E\| \|x\| + O(\varepsilon^2)$$

Now divide through by $\|\vec{x}\|$ $\quad (\vec{x} = \vec{x}(0))$

$$\frac{\|\vec{x}(\varepsilon) - \vec{x}\|}{\|x\|} \leq |\varepsilon| \, \|E\| \, \|A^{-1}\| + O(\varepsilon^2)$$

$$= |\varepsilon| \frac{\|E\|}{\|A\|} \cdot \|A^{-1}\| \cdot \|A\| + O(\varepsilon^2)$$

$$= |\varepsilon| \frac{\|E\|}{\|A\|} K(A) + O(\varepsilon^2)$$

What can we say about

$$\| (A+E)^{-1} - A^{-1} \| \; ?$$

$$A + E = A(I + A^{-1}E) \; ; \quad A^{-1}E = F, \; \|F\| = r < 1$$

$$= A(I - F)$$

$$\| (I-F)^{-1} \| \leq \frac{1}{1 - \|F\|}$$

$$(A+E)^{-1} - A^{-1} = (I + A^{-1}E)^{-1}A^{-1} - A^{-1}$$

$$= (I + A^{-1}E)^{-1}\left(A^{-1} - (I + A^{-1}E)A^{-1}\right)$$

$$= (I + A^{-1}E)^{-1}\left(-A^{-1}EA^{-1}\right)$$

This tells us that

$$\| (A+E)^{-1} - A^{-1} \| \leq \frac{1}{1 - \|A^{-1}E\|} \cdot \|A^{-1}\| \cdot \|E\|$$

$$\frac{\|(A+E)^{-1} - A^{-1}\|}{\|A^{-1}\|} \leq \frac{1}{1-r} \cdot \|A^{-1}\| \cdot \|A\| \cdot \frac{\|E\|}{\|A\|}$$

$\|A^{-1}\| \cdot \|A\| = \kappa(A)$

Say we have a matrix like

$$\begin{pmatrix} 10 & 10 \\ 0 & 10^{-3} \end{pmatrix} + 10^{-4} \begin{pmatrix} 1 & \\ & 1 \end{pmatrix}$$

This only changes the 10^{-3} element substantially

floating Point Arithmetic

Wednesday, October 12, 2005
11:31 AM

Suppose $y = \pm \; .d_1 d_2 \cdots d_n \cdots 10^e$

Clearly not unique, as $.8\overline{9} = .9\overline{0}$

Not even the exponent is unique:

$$0.\overline{9} \cdot 10^0 = 9.1\overline{0} \cdot 10^1$$

$$\tilde{y} = \pm 0.d_1 d_2 - d_s \times 10^e$$

Only s significant digits-this is the "chopped" representation

Or we might do

$$\tilde{y} = \pm \; .d_1 d_2 \cdots \overline{d_s} \times 10^e$$

"rounded" representation where

$$\overline{d_s} = \begin{cases} d_s \text{ if } d_{s+1} < 5 \\ d_s + 1 \bmod 9 \text{ if } d_{s+1} > 5 \end{cases}$$

This could affect all the digits up to s if they're all 9

$\hat{y} = \pm .d_1 \cdots d_s \times \beta^e$

$1 \le d_1 < \beta_1$ call this "normalized"

$0 \le d_j \le \beta$

$m \le e \le M$

if $e > M$, we say "overflow", and if $e < m$, "underflow"

Suppose $s = 1$, with $m = -1$, $M = 1$, $\beta = 10$

Can write

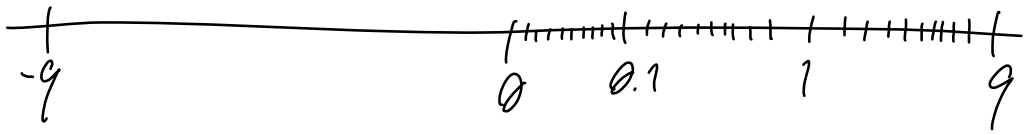
$.1 \times 10^{-1}$ | $.1 \times 10^{0}$ | $.1 \times 10^{1}$

$\vdots$ | $\vdots$ | $\vdots$

$.9 \times 10^{-1}$ | $.9 \times 10^{0}$ | $.9 \times 10^{1}$

$\Rightarrow$ 27 numbers, plus "0"; plus 27 negative numbers

$\Rightarrow$ 55numbers expressible



The point is, the numbers are not uniformly distributed

If we have 2 numbers and add them together, we introduce error; use the notation

$$fl(x+y) = (x+y)(1+\varepsilon)$$

where $|\varepsilon| \leq U_1$ stands for "unit" in the last place

So ε is a function of $x, y,$ and the particular operation, "op"

The IEEE standard says that all computers should handle floating point stuff in the same way

Suppose we have 3 numbers, $x, y, z,$ want $x+y+z$

$$fl(x+y) = (x+y)+z$$

$$fl\left((x+y)(1+\varepsilon) + z\right)$$

$$= (x+y)(1+\varepsilon_1)(1+\varepsilon_2) + z(1+\varepsilon_2)$$

$$= x(1+\varepsilon_1)(1+\varepsilon_2) + y(1+\varepsilon_1)(1+\varepsilon_2) + z(1+\varepsilon_2)$$

So the error is distributed nonuniformly over the summands

Addition Algorithm

Wednesday, October 12, 2005
11:56 AM

$s_n = x_1 + x_2 + \ldots + x_n$
$s_0 = 0$
$s_1 = s_0 + x_1$
$s_2 = s_1 + x_2$
...
$s_n = s_{n-1} + x_n$

$\sigma_0 = 0$

$\sigma_1 = fl(\sigma_0 + x_1)$

$\sigma_n = fl(\sigma_{n-1} + x_n)$

In general

$\sigma_k = fl(\sigma_{k-1} + x_k)$

$= (\sigma_{k-1} + x_k)(1 + \varepsilon_k) \quad |\varepsilon_k| \leq u$

$$\sigma_1 = (\sigma_0 + x_1)(1+\varepsilon_0)$$

$$\sigma_2 = \left[(\sigma_0 + x_1)(1+\varepsilon_0) + x_2\right](1+\varepsilon_1)$$

$$x_1(1+\varepsilon_1) + x_2(1+\varepsilon_1)$$

$$\sigma_3 = x_1(1+\varepsilon_1)(1+\varepsilon_2) + x_2(1+\varepsilon_1)(1+\varepsilon_2) + x_3(1+\varepsilon_2)$$

$$\sigma_k = x_1(1+\delta_1) + x_2(1+\delta_2) + \dots + x_k(1+\delta_k)$$

where $1+\delta_1 = \prod_{j=1}^{n-1}(1+\varepsilon_j)$

$$1+\delta_2 = \prod_{j=2}^{n-1}(1+\varepsilon_j)$$

etc

One way to ameliorate the error is as follows: if they're all positive, add them in increasing order, since smaller #'s have smaller error. Another possibility is to pair them up

x_1
x_2 >

x_3
x_4 >

x_5
x_6 >

The net effect of this is that every # is added into the final # the same number of times

$$\sigma_n = x_1(1+\gamma_1) + x_2(1+\gamma_2) + \cdots + x_n(1+\gamma_n)$$

where $|\gamma_i| \simeq u \log_2 n$

Multiplication

Wednesday, October 12, 2005
12:09 PM

$$p_n = x_1 \cdots x_n$$

$$p_1 = 1 \cdot x_1$$

$$p_2 = p_1 \cdot x_2$$

$$p_3 = p_2 \cdot x_3$$

$$\vdots$$

$$\overline{\Pi}_j = fl\left(x_j \overline{\Pi}_j (1+\delta_j)\right)$$

$$= x_j \overline{\Pi}_j (1+\delta_j)(1+\varepsilon_j)$$

$$\overline{\Pi}_n = x_1 \cdots x_n (1+\varepsilon_1) \cdots (1+\varepsilon_{n-1})$$