Welcome to Ma221! (Feb 10)

Last time: $A = P \cdot L \cdot U$

perm. $\cdot$ $\triangleright$ $\cdot$ $\triangledown$

$L_{ii} = 1$

for $i = 1$ to $n-1$

if $A(i,i) = 0 \neq A(j,i)$ swap rows i & j

record swap in P

$A(i+1:n, i) = A(i+1:n, i) / A(i,i)$

$A(i+1:n, i+1:n) = A(i+1:n, i+1:n)$

$- A(i+1:n, i) \cdot A(i, i+1:n)$

When done: $A =$ 

Show same as traditional GE

for $i = 1$ to $n-1$

... add a multiple of row i to

.... row j to zero out entry (j,i)

... below diagonal

$m = A(j,i) / A(i,i)$

$A(j, i:n) = A(j, i:n) - m \cdot A(i, i:n)$

"Optimize" by

(i) not bother to compute 0 entries:

change to $A(j, i+1:n) = A(j, i+1:n) - m \cdot A(i, i+1:n)$

(2) compute all multipliers m
first, store in "zeroed out" entries

```
for i = 1 to n-1
    for j = i+1:n
        A(j,i) = A(j,i) / A(i,i)
    for j = i+1 to n
        A(j,i+1:n) = A(j,i+1:n) - A(j,i) · A(i,i+1:n)
```

(3) combine loops

```
for i = 1 to n-1
    A(i+1:n, i) = A(i+1:n, i) / A(i,i)
    A(i+1:n, i+1:n) = A(i+1:n, i+1:n)
                      - A(i+1:n, i) · A(i, i+1:n)
```

$$\text{Cost} = \sum_{i=1}^{n-1} 2(n-i)^2 = \frac{2}{3}n^3 + O(n^2)$$

How to pivot, i.e. Choose nonzero $A(i,i)$

Goal: backward stability

$P \cdot L \cdot U = A + E$, $\|E\| = O(\epsilon) \cdot \|A\|$

not guaranteed just by $A(i,i) \neq 0$

Ex: single precision $\epsilon \approx 10^{-7}$

$$A = \begin{bmatrix} 10^{-8} & 1 \\ 1 & 1 \end{bmatrix} \quad A \xrightarrow{\sim} \begin{bmatrix} -1 & 1 \\ 1 & -10^{-8} \end{bmatrix}$$

$\kappa(A) \approx 2.6$ well-conditioned

$\Rightarrow$ expect accurate answer

$$L = \begin{bmatrix} 1 & 0 \\ 10^8 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 10^{-8} & 1 \\ 0 & fl(1 - 10^8 \cdot 1) \end{bmatrix}$$

$$= -10^8$$

$$L \cdot U = \begin{bmatrix} 10^{-8} & 1 \\ 1 & 0 \end{bmatrix}$$

very different from A in (2,2) entry

Get same L, U if $A(2,2)$ were $.5, -1$ etc because $fl(A(2,2) - 10^8 \cdot 1)$ "forgets" $A(2,2)$ if small enough, $O(1)$

So solving $Ax = b$ gives same answer for all these different As, wrong!

Instead: swap rows 1 and 2 $\Rightarrow A(1,1) = 1$ and full accuracy:

Intuition: want large entry of A on diagonal

Recall HW 1.10 : $C = fl(A \cdot B) = A \cdot B + E$

$|E| \le n \cdot \varepsilon \cdot |A| \cdot |B|$

since $A = P \cdot L \cdot U$, get similar bound for GE

Thm (backward error analysis of LU)

if P, L, U from Gauss. El. im

$A - E = P \cdot L \cdot U$

$|E| \le n \cdot \varepsilon \cdot |P| \cdot |L| \cdot |U|$

Cor: Solve $Ax = b$ by GE, forward substition with L, backward with U

The computed $\hat{x}$ satisfies

$(A - F)\hat{x} = b$ where $|F| \le 3n\varepsilon \cdot |P| \cdot |L| \cdot |U|$

Proof of Cor. Assume $P = I$ for simplicity

(imagine running on $P^T A$)

Use HW 1.11

Solve $Ly = b$, get $(L + \delta L)\hat{y} = b$ $|\delta L| \le n \cdot \varepsilon \cdot |L|$

Solve $Ux = \hat{y}$, get $(U + \delta U)\hat{x} = \hat{y}$ $|\delta U| \le n \cdot \varepsilon \cdot |U|$

$b = (L + \delta L)\hat{y} = (L + \delta L)(U + \delta U)\hat{x}$

$= (LU + \delta L \cdot U + L \cdot \delta U + \delta L \cdot \delta U) \cdot \hat{x}$

$= (A - E + \delta L \cdot U + L \cdot \delta U + \delta L \cdot \delta U)\hat{x}$ by Thm

$= (A - F)\hat{x}$

$|F| \le |E| + |\delta L| \cdot |U| + |L| \cdot |\delta U| + |\delta L| \cdot |\delta U|$

$\le n \cdot \varepsilon \cdot |L| \cdot |U|$

$+ n \cdot \varepsilon \cdot |L| \cdot |U|$

$+ n \cdot \varepsilon \, |L| \cdot |U|$

$+ n^2 \varepsilon^2 \, |L| \cdot |U|$

$\cong 3n\varepsilon \, |L| \cdot |U|$ QED of Cor

For backward stability, need

$\|F\| = O(\varepsilon) \cdot \|A\|$

$\|F\| = 3n\varepsilon \, \| \, |L| \cdot |U| \, \|$

want $\| \, |L| \cdot |U| \, \| = O(\|A\|)$

depends on pivot choice

Proof of Thm (recall $P = I$)

Trace through alg: how is $U(i,j)$ computed?

$U(i,j) = A(i,j) - L(i,1) \cdot U(1,j)$

$- L(i,2) \cdot U(2,j)$

$\dots$

$= A(i,j) - \sum_{k=1}^{i-1} L(i,k) \cdot U(k,j)$

Also from taking (i,j) entry of $A = L \cdot U$, solving for $U(i,j)$

$U(i,j)$ = dot product of row i of L col j of U

use previous error analysis for dot products

when $i \geq j$ same idea

$$L(i,j) = \frac{A(i,j) - \sum_{k=1}^{j-1} L(i,k) \cdot U(k,j)}{A(i,i)}$$

again dot product applies QED

$\Rightarrow$ intuition: want $|L(i,j)|$ to be small

(1) Standard: "partial pivoting" (GEPP)

At each step, choose largest entry among $A(i{:}n, i)$

$\Rightarrow |L(k,i)| = |A(k,i)/A(i,i)| \leq 1$

Thm: (easy): with GEPP, $|L| \leq 1$ and $\max(|U(:,i)|) \leq 2^{n-1} \max(|A(:,i)|)$

Bad news: in worst case (attainable) $\|F\| \leq n \geq 2^n \cdot \|A\|$, lose all accuracy in single precision for $n \geq 24$

Good news: hardly ever happens, GEPP is standard alg

Empirical observation is $\frac{\|L\| \cdot \|U\|}{\|A\|} = g \leq n^{2/3}$

If entries of A were random, true with high probability

(2) Complete Pivoting: permute rows and columns, so each $A(i,i)$ is largest in all remaining rows and columns

$P_r^T A P_c^T$ so that $|A(1,1)|$ largest in all of A

$A = P_r \cdot L \cdot U \cdot P_c$

more stable, more expensive than GEPP

$O(n^3)$ more expensive

Thm: $g < n^{(\log n \, / \, 4)}$

Empirically $g < n^{1/2}$

rarely used in practice

(3) Tournament Pivoting: needed to hit communication lower bound $O(n^3/\sqrt{M})$

(4) Threshold Pivoting (sparse case)

Tradeoff stability and sparsity of L, U

Error bound for $Ax = b$

$$\frac{\|x - \hat{x}\|}{\|A\|} \leq \kappa(A) \cdot \text{backward error}$$

$$\leq \kappa(A) \cdot 3 \cdot n \cdot \varepsilon \cdot g$$

g = growth factor

$$\frac{\| |L| \cdot |U| \|}{\|A\|}$$

We can estimate $\kappa(A)$ and g in $O(n^2)$ work

What if error bound too large?
or, error ok, but too slow, so want to use lower precision?

: Try iterative refinement, aka Newton's method

use mixed precision
most work ($O(n^3)$ part) in low prec
rest ($O(n^2)$) in high prec

Low / high could mean

single / double

half / single

bfloat16 / single

double / quad

other combinations (3 precisions)

Do GEPP to solve $Ax=b$ in

low prec, call initial solution $x(1)$

$i=1$

repeat: $r = A \cdot x(i) - b$ in high prec

$O(n^2)$ cost

solve $Ad = r$ in low prec

$O(n^2)$ cost

update $x(i+1) = x(i) - d$ in

low prec, $O(n)$ cost

until "convergence"

we need r in high precision, otherwise

dominated by round off, no information,

and no progress

(still some benefits, both versions

in LAPACK)

Testing "convergence" depends on goals:

(1) Getting a small backward error in high precision

$$\|Ax_{comp} - b\| = O(\varepsilon_{High}) \cdot \|A\| \cdot \|x_{comp}\|$$

or a warning that A too ill-conditioned to converge

Straightforward stopping criterion, since we already compute

$$r = A \cdot x_{comp} - b$$

Motivation: exploit 16-bit accelerators

recent work beyond Newton's method: use GMRES

(2) Getting a small relative error in low precision

$$\frac{\|x_{comp} - x\|}{\|x\|} = O(\varepsilon_{low})$$

or a warning if too ill-conditioned

convergence criterion complicated: avoid being fooled by very ill-conditioned A that look like they are converging

Details in class notes in LAPACK: sgesvxx, dgesvxx