Welcome to Ma221! Lecture 13, Sep 22

4 axes to organize course

1) Math problem: $Ax=b$
2) structure of A: coming up
3) Accuracy

GEPP: Backward stable, rare failures

$$\begin{bmatrix}1&0&0&1\\-1&1&0&1\\-1&-1&1&1\\-1&-1&-1&1\end{bmatrix}\begin{matrix}\\2\\2\;\;4\\2\;\;4\;\;8\end{matrix}$$

repeated doubling can cause instability

"Guaranteed accuracy", rare failures using iterative refinement

convergence proof assumes $\kappa(A)\cdot\varepsilon<1$

4) Fast as possible: minimize communication

Historically, LAPACK reorganized GEPP to use BLAS3 (matmul, GEMM triang solve $LX=B$, TRSM)

Idea similar to induction proof for GEPP, but do b columns at a time, apply updates all at once to trailing matrix

For simplicity, ignore pivoting

$$A = \begin{array}{c|cc} & b & n-b \\ b & A_{11} & A_{12} \\ n-b & A_{21} & A_{22} \end{array} = \left[\begin{array}{c|c} L_{11}\cdot U_{11} & A_{12} \\ \hline L_{21}\cdot U_{11} & A_{22} \end{array}\right]$$

using existing alg to do $\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} = \begin{bmatrix} L_{11} \\ L_{21} \end{bmatrix} U_{11}$

$$= \left[\begin{array}{c|c} L_{11}\cdot U_{11} & L_{11}\cdot U_{12} \\ \hline L_{21}\cdot U_{11} & A_{22} \end{array}\right]$$

where we solved $A_{12} = L_{11}\cdot U_{12}$ for U_{12} using TRSM

$$= \left[\begin{array}{c|c} L_{11} & 0 \\ \hline L_{21} & I \end{array}\right] \left[\begin{array}{c|c} U_{11} & U_{12} \\ \hline 0 & \underbrace{A_{22} - L_{21}\cdot U_{12}}_{\text{Schur complement} = S} \end{array}\right]$$

mat mul, GEMM

repeat on S

Often very fast, but for some combinations of n and cache size M can't choose b to minimize comm i.e. reach $O\left(\frac{n^3}{\sqrt{M}}\right)$ words moved

Just as for matmul, there is a recursive cache oblivious alg that reaches $O\left(\frac{n^3}{\sqrt{M}}\right)$ (Toledo, 1997)

High Level alg:

DO LU on left half of A

Update right half

(U at top, Schur complement at bottom)

Do LU on Schur Complement

function $[L, U] = RLU(A)$ ... Recursive LU

... assume A $n \times m$, $n \geq m$, m power of 2

if $m = 1$ ... one column

pivot so A_{11} largest entry, pivot rest of matrix

$L = A / A_{11}$, $U = A_{11}$

else ... write $A = \begin{array}{cc} & \begin{array}{cc} m/2 & m/2 \end{array} \\ \begin{array}{c} \frac{m}{2} \\ n-\frac{m}{2} \end{array} & \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \end{array}$ $\quad L_1 = \begin{array}{cc} m/2 & \\ \begin{bmatrix} L_{11} \\ L_{12} \end{bmatrix} & \begin{array}{c} m/2 \\ n-m/2 \end{array} \end{array}$

$[L_1, U_1] = RLU\left(\begin{bmatrix} A_{11} \\ A_{22} \end{bmatrix}\right)$ ... LU of left half

Solve $A_{12} = L_{11} \cdot U_{12}$ for U_{12} ... update U

$A_{22} = A_{22} - L_{21} \cdot U_{12}$ ... update Schur compl.

$[L_2, U_2] = RLU(A_{22})$

$L = \left[L_1, \begin{bmatrix} 0 \\ L_2 \end{bmatrix}\right]^{n \times m}$, $U = \begin{bmatrix} U_1 & U_{12} \\ & U_2 \end{bmatrix}^{m \times m}$

correctness by induction

Cost: Recurrences (for $m = n$)

$A(n) = \#$ arith ops $= \frac{2}{3} n^3 + O(n^2)$

same flops as usual alg

similar recurrence to matmul

$$W(n) = \#\text{ word moved} = O\left(\frac{n^3}{\sqrt{M}}\right)$$

RLU: only hits lower bound for #words moved

not # messages

To minimize # messages

① Replace partial pivoting by
tournament pivoting (see notes)

② Keep partial pivoting, more complicated
datastructure; payoff unclear

How to use Strassen-like algs?

Can modify RLU to run in $O(n^\omega)$ flops
if matmul does

① multiply $L_{21} \cdot U_{12}$ using $O(n^\omega)$ matmul

② solve $A_{12} = L_{11} \cdot U_{12}$ by divide & conquer
invert L_{11} (not as stable as GEPP)

$$T^{-1} = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}^{-1} = \begin{bmatrix} T_{11}^{-1} & -T_{11}^{-1} \cdot T_{12} \cdot T_{22}^{-1} \\ 0 & T_{22}^{-1} \end{bmatrix}$$ use $O(n^\omega)$ matmul

recall $L_{ii} = 1$, $|L_{ij}| \le 1$ so
should be reasonably conditioned

Where to find implementations

Matlab: $A\backslash b$, or $[L,U,P] = lu(A)$

rcond, condest to estimate $\kappa(A)$

LAPACK: xGETRF : GEPP x = S/D/C/Z

xGETRF2: GEPP recursively

xGESV solve $Ax=b$

xGESVX iterative refinement in precison x

xGESVXX: iterative refinement residual in double precision of x

xGECON for condition est

many other libraries

ScaLapack, SLATE cluster

PLASMA multicore

MAGMA GPUs

⋮

Exploiting Structure in A

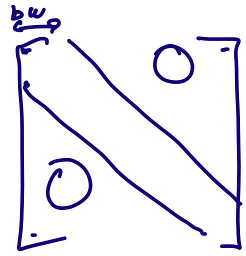
A symmetric positive definite: (spd)

- Cholesky, no pivoting, $\frac{1}{2}$ flops of GEPP

A symmetric only

- save half flops, pivoting required but care needed to keep symmetry

A: band matrix
bw = bandwidth



cost in flops drops from $O(n^3)$ to $(bw^2 \cdot n)$
space drops from $O(n^2)$ to $O(bw \cdot n)$

A: sparse matrix: lots of zeros
cost, space drop significantly
depends on pattern of nonzeros
many complicated algs, lots of software

A: structured matrices: dense but
depend on $O(n)$ parameters
Vandermonde: $V(i,j) = x_i^{j-1}$
Toeplitz: $T(i,j) = t_{i-j}$
many more, discuss most common

Symmetric (Hermitian) Positive Definite
spd or hpd for short

Def: A real and spd iff $A = A^T$
and $x^T A x > 0 \quad \forall x \neq 0 \quad x$ real vector
A complex and hpd iff $A = A^H$
and $x^H A x > 0 \quad \forall x \neq 0 \quad x$ complex vector

Lemma: (just real case)

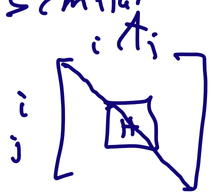
1) X nonsingular $\Rightarrow$ A s.p.d. iff $X^T A X$ s.p.d.

pf: A s.p.d. and $x \neq 0 \Rightarrow Xx \neq 0 \Rightarrow$

$$0 \neq (Xx)^T A (Xx) = x^T X^T A X x$$
$$= x^T (X^T A X) x$$

$\Rightarrow X^T A X$ s.p.d. other direction similar

2) A s.p.d. $H = A(j:k, j:k)$



H "principal submatrix"

H s.p.d.

pf: A spd. $y \neq 0 \Rightarrow 0 \neq x = \begin{bmatrix} 0 \\ y \\ 0 \end{bmatrix}$

$\Rightarrow 0 \neq x^T A x = y^T H y \Rightarrow H$ s.p.d

3) A spd iff $A = A^T$ and all evals $\lambda_i > 0$