NLA 2021-2022
Linear equation solving

Martin Sombra

22 September 2021

Linear equation solving

We want to solve

$$A\,x = b$$

for a nonsingular $n \times n$ matrix A and an n-vector b

Gaussian elimination consists in computing a factorization

$$A = P\,L\,U$$

with P a permutation, L a unit lower triangular, and U an upper triangular

Allows to solve $A\,x = b$ by solving the simpler equations

1. $P\,z = b$ (permute the entries of b)
2. $L\,y = z$ (forward substitution)
3. $U\,x = y$ (backwards substitution)

Constructing the PLU factorization

Set

$$A = \begin{bmatrix} a_{1,1} & \dots & a_{1,n} \\ \vdots & & \vdots \\ a_{n,1} & \dots & a_{n,n} \end{bmatrix}.$$

For $n = 1$

$$P = L = \begin{bmatrix} 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} a_{1,1} \end{bmatrix}$$

Let $n \geqslant 2$ and choose k such that $a_{k,1} \neq 0$

Gaussian elimination with partial pivoting (GEPP)

k such that $|a_{k,1}|$ is maximal

Swap rows 1 and k premultiplying by the permutation matrix

$$P_1 = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & & & 0 & & & \\ \vdots & & \ddots & & \vdots & & & \\ 0 & & & 1 & 0 & & & \\ 1 & 0 & \cdots & 0 & 0 & & & \\ 0 & & & & & 1 & & \\ \vdots & & & & & & \ddots & \\ 0 & & & & & & & 1 \end{bmatrix}.$$

Consider the 2×2-block

$$P_1^T A = \begin{bmatrix} a_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

Constructing the PLU factorization (cont.)

Then

$$\begin{bmatrix} a_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} = \begin{bmatrix} 1 & \mathbb{0} \\ L_{2,1} & \mathbb{1}_{n-1} \end{bmatrix} \begin{bmatrix} u_{1,1} & U_{1,2} \\ \mathbb{0} & \widetilde{A}_{2,2} \end{bmatrix}$$

with

$$u_1 = a_{1,1}, \quad L_{2,1} = a_{1,1}^{-1} A_{2,1}, \quad U_{1,2} = A_{1,2} \quad \text{and} \quad \widetilde{A}_{2,2} = A_{2,2} - L_{2,1}\, U_{1,2}$$

The $(n-1) \times (n-1)$ matrix $\widetilde{A}_{2,2}$ is the *Schur complement*

By the inductive hypothesis (case $n-1$):

$$\widetilde{A}_{2,2} = \widetilde{P}\, \widetilde{L}\, \widetilde{U}$$

Constructing the PLU factorization (cont.)

Then

$$P_1^T A = \begin{bmatrix} 1 & \mathbb{0} \\ L_{2,1} & \mathbb{1}_{n-1} \end{bmatrix} \begin{bmatrix} u_{1,1} & U_{1,2} \\ \mathbb{0} & \widetilde{P}\,\widetilde{L}\,\widetilde{U} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \mathbb{0} \\ \mathbb{0} & \widetilde{P} \end{bmatrix} \begin{bmatrix} 1 & \mathbb{0} \\ \widetilde{P}^T L_{2,1} & \widetilde{L} \end{bmatrix} \begin{bmatrix} u_{1,1} & U_{1,2} \\ \mathbb{0} & \widetilde{U} \end{bmatrix}$$

Hence $A = P\,L\,U$ with

$$P = P_1 \begin{bmatrix} 1 & \mathbb{0} \\ \mathbb{0} & \widetilde{P} \end{bmatrix}, \quad L = \begin{bmatrix} 1 & \mathbb{0} \\ \widetilde{P}^T L_{2,1} & \widetilde{L} \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} u_{1,1} & U_{1,2} \\ \mathbb{0} & \widetilde{U} \end{bmatrix}.$$

GEPP $\rightsquigarrow$ $L = [l_{i,j}]_{i,j}$ with $|l_{i,j}| \leqslant 1$ for all i, j

Complete pivoting

Gaussian elimination with complete pivoting (GECP): takes $|a_{k,l}|$ maximal among all the entries
$\leadsto$ swaps the rows 1 and k, and the columns 1 and l

Gives a factorization

$$A = P_1\, L\, U\, P_2$$

with P_1 and P_2 permutations

Can be more *numerically stable* but is also *more expensive* (in terms of speed)

The GEPP algorithm

for $i = 1, \ldots, n-1$

swap row k and row i of A and L for k such that $|a_{k,i}| = \max_{i \leqslant p \leqslant n} |a_{p,i}|$

for $j = i+1, \ldots, n$ (compute column i of L)

$l_{j,i} \leftarrow \frac{a_{j,i}}{a_{i,i}}$

for $j = 1, \ldots, n$ (compute row i of U)

$u_{i,j} \leftarrow a_{i,j}$

for $j, k = i+1, \ldots, n$ (update $A_{2,2}$)

$u_{j,k} \leftarrow a_{j,k} - l_{j,i}\, u_{i,k}$

Modified GEPP

Useful remark: the i-th column of A only used to compute the i-th column of L, and the i-th row of A only used to compute the i-th row of U

for $i = 1, \ldots, n-1$

swap row k and row i of A and L for k such that $|a_{k,i}| = \max_{i \leqslant p \leqslant n} |a_{p,i}|$

for $j = i+1, \ldots, n$

$\cancel{l_{j,i}} \leftarrow \frac{a_{j,i}}{a_{i,i}}$, (replace by $a_{j,i}$)

$\cancel{\text{for } j = 1, \ldots, n}$

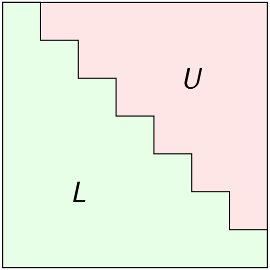
$\cancel{u_{i,j} \leftarrow a_{i,j}}$

for $j, k = i+1, \ldots, n$

$\cancel{u_{j,k}} \leftarrow a_{j,k} - \cancel{l_{j,i}}\, \cancel{u_{i,k}}$ (replace by $a_{j,k}$, $a_{j,i}$ and $a_{i,k}$)

Modified GEPP (cont.)

We need no extra space to store L and U!



Complexity

Recall the asymptotic formulae for $k \geqslant 1$:

$$\sum_{i=1}^{n} i^k = \frac{n^{k+1}}{k+1} + O(n^k),$$

The complexity (# flops) of GEPP on $n \times n$ matrices is

$$\sum_{i=1}^{n-1} \left(\sum_{j=i+1}^{n} 1 + \sum_{j=i+1}^{n} \sum_{k=i+1}^{n} 2 \right) = \sum_{i=1}^{n-1} ((n-i) + 2\,(n-i)^2) = \frac{2}{3}\,n^3 + O(n^2)$$

Forward and backward substitution have each a complexity of

$$n^2 + O(n)$$

Hence GEPP solves the equation $A\,x = b$ with complexity

$$\frac{2}{3}\,n^3 + O(n^2)$$

Floating point arithmetic

A *floating point number* is

$$f = \pm 0.d_1 d_2 \ldots d_t \times \beta^e$$

$$\beta \geqslant 2, \quad 0 \leqslant d_i < \beta \text{ with } d_1 \neq 0 \quad \text{and} \quad L \leqslant e \leqslant U$$

$\beta \geqslant 2$ the *base*, L the *underflow* and U the *overflow*

The *range* is

$$\beta^{L-1} \leqslant |f| \leqslant \beta^U (1 - \beta^{-t}).$$

Floating point operations are defined by picking the closest floating point number:

$$a \odot b \coloneqq \mathrm{fl}(a \cdot b)$$

Machine epsilon: a bound for the relative error of roundoffs:

$$\frac{|a - \mathrm{fl}(a)|}{|a|} \leqslant \varepsilon = \frac{\beta^{1-t}}{2}$$

Single precision IEEE standard

s	e	q
1	8	23
sign	exponent	coefficient/mantissa

The *coded number* is

$$f = (-1)^s(1+q)\,2^{e-127}$$

Corresponds to $t = 24$, $\beta = 2$, $L = -126$ and $U = 127$

The maximal *relative error* is

$$2^{-24} \approx 6 \cdot 10^{-8}$$

and the *range* is

$$2^{-127} \approx 6 \cdot 10^{-38} \leqslant f \leqslant 2^{129}(1 - 2^{-24}) \approx 7 \cdot 10^{39}$$

IEEE double precision standard

s	e	q
1	11	52
sign	exponent	coefficient/mantissa

The coded number is

$$f = (-1)^s (1+q)\, 2^{e-1023}.$$

Corresponds to $t = 53$, $\beta = 2$, $L = -1022$ and $U = 1026$

The maximal error is

$$2^{-53} \approx 6 \cdot 10^{-16}$$

and the range is

$$2^{-1022} \approx 2 \cdot 10^{308} \leqslant f \leqslant 2^{1029}(1 - 2^{-24}) \approx 7 \cdot 10^{309}$$

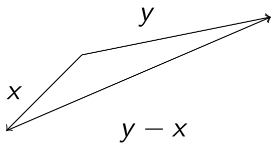
Vector and matrix norms

A *norm* on $\mathbb{R}^n$ is a function

$$\|\cdot\|: \mathbb{R}^n \longrightarrow \mathbb{R}$$

such that

1. for every vector x we have that $\|x\| \geqslant 0$, and $\|x\| = 0$ if and only if $x = 0$
2. for every vector x and scalar α we have that $\|\alpha\, x\| = |\alpha|\, \|x\|$
3. for every vectors x, y we have that $\|x + y\| \leqslant \|x\| + \|y\|$ *(triangle inequality)*.



Example: for $1 \leqslant p \leqslant +\infty$ the *p-norm* is defined as

$$\|x\|_p = \begin{cases} \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} & \text{if } 1 \leqslant p < +\infty \\ \max_i |x_i| & \text{if } p = +\infty \end{cases}$$

Any two norms can be compared: there are $c_1, c_2 > 0$ such that

$$\|\cdot\|_1 \leqslant c_1 \|\cdot\|_2 \quad \text{and} \quad \|\cdot\|_2 \leqslant c_2 \|\cdot\|_1$$

Example:

$$\|\cdot\|_2 \leqslant \|\cdot\|_1 \leqslant n^{1/2} \|\cdot\|_2 \quad \text{and} \quad \|\cdot\|_{infty} \leqslant \|\cdot\|_1 \leqslant n \|\cdot\|_\infty$$

Vector and matrix norms (cont.)

A *matrix norm* is a norm on $\mathbb{R}^{n\times n}$ s.t. for all A, B we have that

$$\|A\,B\| \leqslant \|A\|\,\|B\|$$

Example: the *Frobenius norm*

$$\|A\|_{\mathrm{F}} = \left(\sum_{i,j} |a_{i,j}|^2\right)^{1/2}$$

Vector and matrix norms (cont.)

Definition: Given a norm on $\mathbb{R}^n$, the associated *operator norm* is

$$\|A\| = \sup_{x \neq 0} \frac{\|A\,x\|}{\|x\|}$$

It is a matrix norm

Properties:

1. $\|A\|_\infty = \max_i \sum_j |a_{i,j}|$ *(maximum row sum)*
2. $\|A\|_1 = \max_j \sum_i |a_{i,j}|$ *(maximum column sum)*
3. $\|A\|_2 = \lambda_{\max}(A^* A)^{1/2}$ with $A^* = \overline{A}^T$ and $\lambda_{\max}$ the maximal eigenvalue
4. if Q and Q' are *orthogonal* then

$$\|Q\,A\,Q'\|_2 = \|A\|_2 \quad \text{and} \quad \|Q\,A\,Q'\|_{\mathrm{F}} = \|A\|_{\mathrm{F}}$$

In particular $\|Q\|_2 = 1$ and $\|Q\|_{\mathrm{F}} = n^{1/2}$

Perturbation theory

A matrix A is *well/badly (or ill) conditionned* if small changes in A can cause small/large changes in the solution of

$$A\,x = b$$

Let x and $\widehat{x} = x + \delta x$ solutions to

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

We have that

$$\begin{array}{rcc} & (A+\delta A)\,(x+\delta x) & = b - \delta b \\ - & A\,x & = \quad b \\ \hline & \delta\,A\,x + (A + \delta\,A)\,\delta x = & \delta b \end{array}$$

Then

$$\delta x = A^{-1}(-\delta A\,\widehat{x} + \delta b)$$

The condition number

Fix a norm $\|\cdot\|$. Then

$$\|\delta x\| \leqslant \|A^{-1}\| \left(\|\delta A\| \, \|\widehat{x}\| + \|\delta x\| \right)$$

or equivalently

$$\frac{\|\delta x\|}{\|\widehat{x}\|} \leqslant \kappa_{\|\cdot\|} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \, \|\widehat{x}\|} \right) \tag{1}$$

with

$$\kappa_{\|\cdot\|} := \|A\| \, \|A^{-1}\|$$

the *condition number* of A with respect to $\|\cdot\|$

Precision of approximations

Let $\lambda \in \mathbb{R}$ and $\widehat{\lambda} = \lambda + \delta\lambda$ an approximation of λ with k correct digits in base $\beta \geqslant 2$. Then

$$\lambda = \beta^e \times d_1 \cdots d_k d_{k+1} \cdots \quad \text{and} \quad \widehat{\lambda} = \beta^e \times d_1 \cdots d_k \widetilde{d}_{k+1} \cdots$$

and so

$$\frac{|\delta\lambda|}{|\lambda|} \leqslant \beta^{-k}$$

or equivalently

$$-\log_\beta \left(\frac{|\delta\lambda|}{|\lambda|} \right) \geqslant k.$$

Roundoff with IEEE single precision and double precision give approximations with 24 and 53 correct bits, respectively:

$$-\log_2 \left(\frac{|\delta_{\text{single}}\lambda|}{|\lambda|} \right) \geqslant 24 \quad \text{and} \quad -\log_2 \left(\frac{|\delta_{\text{double}}\lambda|}{|\lambda|} \right) \geqslant 53.$$

Lost of precision

The inequality (1) translates into

$$-\log_\beta \frac{\|\delta x\|}{\|x\|} \geqslant -\log_\beta \kappa(A) - \log_\beta \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \, \|\widehat{x}\|} \right)$$

Warning: ill-conditionned matrices matrices destroy the quality of your approximations!

For instance, for *exact* data truncated with IEEE single or double precision, the computed result of $A\,x = b$ will be meaningless as as soon

$$\kappa(A) > 2^{24} \approx 6 \cdot 10^8 \text{ (single precision)}$$

and

$$\kappa(A) > 2^{53} \approx 10^{16} \text{ (double precision)}$$

Distance to the ill-posed problems

The condition number of the 2-norm has a geometrical interpretation as the inverse if its distance to the set of ill-posed problems:

$$\kappa_2(A) = \frac{1}{\text{distance}(A, \Sigma)} \tag{2}$$

with $\Sigma = \{A \mid \det(A) = 0\}$