NLA 2021-2022
Linear equation solving (part 2)

Martin Sombra

29 September 2021

Perturbation theory in linear equation solving

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

$$Ax = b$$

Perturbation theory in linear equation solving

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

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

We have that

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

Then

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

The condition number

Fix a norm $\|\cdot\|$

Then

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

or equivalently

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

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:

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

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:

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

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

Error analysis in GEPP

We want to apply the two steps:

1. analyse roundoff errors to show that the matrix

$$\widehat{A}_{\mathrm{GEPP}} := P_{\mathrm{GEPP}} L_{\mathrm{GEPP}} U_{\mathrm{GEPP}}$$

   has a small relative error *(backward analysis)*

2. apply perturbation theory to bound the error in the computed solution x_{GEPP} of the equation

$$A_{\mathrm{GEPP}}\, x = b$$

Error analysis in GEPP (cont.)

Rounding off the entries of A gives $\widehat{A} = A + \delta A$ with

$$\frac{\|\delta A\|}{\|A\|} < \varepsilon \qquad \text{(machine epsilon)}$$

By perturbation theory, this error will be amplified to

$$\frac{\|\delta x\|}{\|x\|} < \kappa_{\|\cdot\|}(A)\,\varepsilon.$$

To keep this bound, for $\delta_{\mathrm{GEPP}} A \coloneqq A_{\mathrm{GEPP}} - A$ we want

$$\frac{\|\delta_{\mathrm{GEPP}} A\|}{\|A\|} \leqslant C\,\varepsilon$$

with C as small as possible

The need of pivoting

Apply LU factorization without pivoting to the matrix

$$A = \begin{bmatrix} \eta & 1 \\ 1 & 1 \end{bmatrix}$$

with η a power of the base β that is smaller than ε, so that

$$1 \oplus \eta = \mathrm{fl}(1 + \eta) = 1$$

For instance

$$\beta = 10, \quad \varepsilon = 0.5 \cdot 10^{-3} \quad \text{and} \quad \eta = 10^{-4}$$

The need of pivoting (cont.)

Set

$$A = L\,U = \begin{bmatrix} 1 & 0 \\ \eta^{-1} & 1 \end{bmatrix} \begin{bmatrix} \eta & 1 \\ 0 & 1-\eta^{-1} \end{bmatrix}$$

Then

$$L_{\mathrm{GEWP}} = \begin{bmatrix} 1 & 0 \\ \eta^{-1} & 1 \end{bmatrix} \quad \text{and} \quad U_{\mathrm{GEWP}} = \begin{bmatrix} \eta & 1 \\ 0 & -\eta^{-1} \end{bmatrix}$$

and so

$$A_{\mathrm{GEWP}} = L_{\mathrm{GEWP}}\,U_{\mathrm{GEWP}} = \begin{bmatrix} \eta & 1 \\ 1 & 0 \end{bmatrix},$$

is *not* close to A!

$$\frac{\|\delta A_{\mathrm{GEWP}}\|_\infty}{\|A\|_\infty} = \frac{\left\| \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \right\|_\infty}{\left\| \begin{bmatrix} \eta & 1 \\ 1 & 1 \end{bmatrix} \right\|_\infty} = \frac{1}{2}$$

$\rightsquigarrow$ GEWP is not backward stable

The need of pivoting (cont.)

The solution of $A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ is $x \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Solving

$$L_{\text{GEWP}} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

gives $y_1 = 1$ and $y_2 = 2 \ominus \eta^{-1} = -\eta^{-1}$. Then

$$U_{\text{GEWP}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -\eta^{-1} \end{bmatrix}$$

gives $x_2 = \frac{-\eta^{-1}}{-\eta^{-1}} = 1$ and $x_1 = \frac{1 \ominus 1}{1 \ominus \eta} = 0$. Hence

$$x_{\text{GEWP}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

is *not* close to x

The need of pivoting (cont.)

The instability is also reflected in the conditions numbers:

$$\|A\|_\infty \approx 4 \qquad \text{well-conditioned}$$

whereas

$$\|L\|_\infty, \|U\|_\infty \approx \eta^{-2} \qquad \text{ill-conditioned}$$

Formal error analysis of GEPP

When the intermediate quantities are too large, the information in A can be easily lost

Suppose that A is already pivoted. Then

$$A = L_{\mathrm{GEPP}}\, U_{\mathrm{GEPP}} + E \quad \text{with } |E| \leqslant n\,\varepsilon\,|L|\,|U|$$

where

- $|E|$ the $n \times n$ matrix whose entries are the absolute values of those of E (and similarly for $|L|$ and $|U|$)
- ε the machine epsilon

Formal error analysis of GEPP (cont.)

Hence

$$\|A - A_{\mathrm{GEPP}}\|_\infty \leqslant n\,\varepsilon\,\||L|\|_\infty \||U|\|_\infty \leqslant n^3\,\varepsilon\,g_{\mathrm{GEPP}}\,\|A\|_\infty$$

where

$$g_{\mathrm{GEPP}} = \frac{\max_{i,j}|u_{i,j}|}{\max_{i,j}|a_{i,j}|} \quad \text{the pivot growth}$$

because

- $|l_{i,j}| \leqslant 1$ and so $\|L\|_\infty \leqslant n$
- $|u_{i,j}| \leqslant g_{\mathrm{GEPP}}\|A\|_\infty$ and so $\|U\|_\infty \leqslant n\,g_{\mathrm{GEPP}}\,\|A\|_\infty$

Thus

$$\frac{\|\delta_{\mathrm{GEPP}}A\|_\infty}{\|A\|_\infty} \leqslant n^3 \varepsilon\, g_{\mathrm{GEPP}} \tag{3}$$

Formal error analysis of GEPP (cont.)

In general $g_{\mathrm{GEPP}} \leqslant 2^{n-1}$, and this bound can be attained:

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & -1 & 1 & 0 \\ -1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{bmatrix}$$

Formal error analysis of GEPP (cont.)

This bound in (3) is too pesimistic in practice, since typically

$$\| |L| \|_\infty \, \| |U| \|_\infty \approx \|A\|_\infty$$

If this is the case, then

$$\frac{\|\delta_{\mathrm{GEPP}} A\|}{\|A\|} \lesssim n\,\varepsilon$$

and GEPP would be stable

We thus say that GEPP is “backward stable in practice” (!?)