Lecture 20 (hrs. 38,39,40) - November 11, 2025, 15:30 - 18:30 F3

(2.41) Example (part two).

Let us assume that the elastic constants c_k are known with uncertainty. For example, let us assume that, for k = 1, 2, 3, we have:

$$c_k' = c_k + \delta c_k \qquad \text{where} \qquad |\delta c_k| < 1 \text{ N/m}$$

The system A z = b is transformed into the *perturbed system* (A + δA) z = b + δb with:

$$\delta A = \begin{bmatrix} \delta c_1 + \delta c_2 & -\delta c_2 \\ -\delta c_2 & \delta c_2 + \delta c_3 \end{bmatrix} \quad , \quad \delta b = \begin{bmatrix} 0 \\ h\ \delta c_3 \end{bmatrix}$$

The relative measure of the data perturbations satisfy:

$$\varepsilon_A = N_1(\delta A)/N_1(A) < 10^{-2} \quad , \quad \varepsilon_b = N_1(\delta b)/N_1(b) < 10^{-2}$$

Moreover:

$$c_1(A)\ \varepsilon_A \leqslant 3 \times 10^{-2}$$

According to the Conditioning Theorem, for the deviation of the solution $\hat{z}$ of the perturbed system from the solution z^* we have the following limitation:

$$\varepsilon_x \leqslant \frac{c_1(A)}{1 - c_1(A)\,\varepsilon_A}(\varepsilon_A + \varepsilon_b) \approx 6.2 \times 10^{-2}$$

As regards the deviation of the components, this time we have:

$$\varepsilon_{x,1} \leqslant 0.19\ (19\,\%) \quad , \quad \varepsilon_{x,2} \leqslant 0.09\ (9\,\%)$$

(2.42) Example (part three).

Let $\hat{z}$ now be a column (for example obtained by the calculator using a procedure for solving the system A z = b) to be used as an *approximation* of z^*. To obtain a bound on the error committed, we proceed as in Remark (2.39) of Lecture 19.

The residual vector is:

$$r = A\,\hat{z} - b$$

(1) We read $\hat{z}$ as the solution of the perturbed system A z = b + r. By the Conditioning Theorem:

$$\frac{N_1(\hat{z} - z^*)}{N_1(z^*)} \leqslant c_1(A)\ \frac{N_1(r)}{N_1(b)}$$

Question: Are there perturbations of the parameters δg, δc_k, δm_k, δh that generate perturbations of the data $\delta A = 0$ and $\delta b = r$ (i.e., is it possible to give a 'physical meaning' to the perturbed system A z = b + r) ?

Answer: Yes. For example: $\delta g = 0$, $\delta c_k = 0$, $\delta h = 0$ and $\delta m_1 = r_1/g$, $\delta m_2 = r_2/g$.

(2) We seek $M \in R^{2\times 2}$ such that $M\hat{z} = -r$, and we read $\hat{z}$ as a solution of the perturbed system $(A + M)\,z = b$. By the Conditioning Theorem, defined $\varepsilon_A = \|M\|_1/\|A\|_1$ it is:

$$\text{if} \quad c_1(A)\varepsilon_A \leqslant 1 \text{ then } \frac{N_1(\hat{z}-z^*)}{N_1(z^*)} \leqslant \frac{c_1(A)\varepsilon_A}{1-c_1(A)\varepsilon_A}$$

Question: Are there perturbations of the parameters δg, δc_k, δm_k, δh that generate perturbations of the data $\delta A = M$ and $\delta b = 0$ (i.e., is it possible to give a 'physical meaning' to the perturbed system $A\,z = b + r$) ?

(2.43) Example.

Let:

$$\hat{z} = \begin{bmatrix} 1.8 \\ 3.4 \end{bmatrix} m$$

Then:

$$r = A\,\hat{z} - b = \begin{bmatrix} 10.19 \\ -9.81 \end{bmatrix} N$$

We seek for α and β such that, defined:

$$M = \begin{bmatrix} \alpha+\beta & -\beta \\ -\beta & \beta \end{bmatrix}$$

we get:

$$M\hat{z} = -r$$

We obtain a system of two equations in the unknowns α and β whose only solution is:

$$\alpha = -(r_1 + r_2)/\hat{z}_1 \approx -0.21 \text{ N/m} \qquad \text{and} \qquad \beta = -r_2/(\hat{z}_2 - \hat{z}_1) \approx -6.13 \text{ N/m}$$

Then:

$$\varepsilon_A \approx 4.1 \times 10^{-2} \quad \text{and} \quad c_1(A)\,\varepsilon_A \approx 0.12 < 1$$

hence, by the Conditioning Theorem:

$$\varepsilon_x \leqslant 0.14 \text{ (approximately)}$$

Finally, the answer is yes.: $\delta g = 0$, $\delta m_k = 0$, $\delta h = 0$ and $\delta c_1 = \alpha$ N/m, $\delta c_2 = \beta$ N/m, $\delta c_3 = 0$.

(2.2) STUDY OF A LINEAR SYSTEM IN F(β,m)

(2.44) Remark (study using EGP).

Let $A \in R^{n\times n}$ and $b \in R^n$. The procedure for studying the system $A\,x = b$ using the EGP

procedure is:

```
(S,D,P) = EGP(A);                                                    in R
      if there exists k s.t. d_kk = 0 then STOP;
        otherwise
            c = SA(S,Pb);
            x* = SI(D,c)
```

When using a computer, whose set of machine numbers is $F(\beta,m)$, the procedure turns into:

```
(Ŝ,D̂,P̂) = EGP_M(Â);                                                  in F(β,m)
      if there exists k s.t. d̂_kk = 0 then STOP;
        otherwise
            ĉ = SA_M(S,Pb̂);
            x̂ = SI_M(D,c)
```

where:

- EGP_M, SA_M and SI_M are, respectively, the EGP, SA and SI procedures in which each arithmetic operation is replaced by the corresponding predefined function,
- $\hat{A}$ and $\hat{b}$ are, respectively, the matrix rd(A) and the column rd(b) whose elements are the rounded in $F(\beta,m)$ of the corresponding elements of A and b.

(2.45) Example.

Recalling Theorem (1.38) of Lecture 6, for each component of the matrix $\hat{A} = rd(A)$ and of the column $\hat{b} = rd(b)$ we have:

$$\hat{a}_{ij} = rd(a_{ij}) = (1 + \varepsilon_{ij})\, a_{ij} \qquad , \qquad \hat{b}_i = rd(b_i) = (1 + \varepsilon_i)\, b_i$$

where $|\varepsilon_{ij}| \leqslant u$ and $|\varepsilon_i| \leqslant u$ for every i and j. It follows that, using for example the norm one in R^n, the absolute measures of the perturbations satisfy:

$$\| \delta A \|_1 \leqslant u \| A \|_1 \qquad , \qquad N_1(\delta b) \leqslant u\, N_1(\delta b)$$

hence the relative measures satisfy:

$$\varepsilon_A \leqslant u \qquad e \qquad \varepsilon_b \leqslant u$$

If $c_1(A)\, u \leqslant 1$ then $c_1(A)\, \varepsilon_A \leqslant 1$ and, by the Conditioning Theorem (Theorem (2.36) of Lecture 18) it is:

$$\varepsilon_x \leqslant 2\, \frac{c_1(A)\, u}{1 - c_1(A)\, u} \equiv \Lambda$$

When the calculator *reads the data* A and b, it changes them (except when the data components are in $F(\beta,m)$) and the system $A\,x = b$ is replaced by the system $\hat{A}\,x = \hat{b}$. This substitution, in the best possible case where the effect of the substitutions of EGP, SA and SI with EGP_M, SA_M and SI_M is negligible, *can generate* a deviation of the solution x^* of relative size Λ. Therefore, in the usual case where the effect of the substitutions of EGP, SA and SI with EGP_M, SA_M and SI_M is not negligible, *it is not reasonable* to expect a

deviation between x^* and the approximation $\hat{x}$ obtained by the calculator *smaller than* Λ.

(2.46) Example.

Consider the following 'almost-ideal' situation:

- $\hat{A} = A$, $\hat{b} = b$ - the components of the data are in $F(\beta,m)$;
- $EGP_M(A) = EGP(A) = (S,D,P)$ - the result of EGP_M 'is exact', and D is invertible;
- $SA_M(S,Pb) = \hat{c} = rd(c)$ - the result of SA_M is 'almost ideal';
- $SI_M(D,\hat{c}) = SI(D,\hat{c})$ - the result of SI_M 'is exact'.

Under these assumptions we have: $x^* = SI(D,c)$ is the solution of the system $D\,x = c$, $\hat{x} = SI(D,\hat{c})$ is the solution of the system $D\,x = \hat{c}$. Introducing the perturbation $\delta c = \hat{c} - c$ we have, using the norm N_1 (see the previous example):

$$N_1(\delta c) \leqslant u\, N_1(c) \qquad \text{hence} \qquad \varepsilon_c \leqslant u$$

By the Conditioning Theorem we have:

$$\varepsilon_x \leqslant c_1(D)\, \varepsilon_c \leqslant c_1(D)\, u$$

The limitation of the relative measure of the deviation depends on $c_1(D)$. Let us rewrite:

$$c_1(D) = c_1(A)\, \frac{c_1(D)}{c_1(A)}$$

The limitation depends on the *amplification factor of the condition number* $c_1(D)/c_1(A)$.

(2.47) Example.

Let $\gamma \in (0,1)$ and $A = \begin{bmatrix} \gamma & 1 \\ 1 & 0 \end{bmatrix}$. It is:

- $\| A \|_1 = 1 + \gamma < 2$
- $A^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & -\gamma \end{bmatrix}$ and hence $\| A^{-1} \|_1 = 1 + \gamma$ and $c_1(A) = (1 + \gamma)^2 < 4$
- $EGP(A) = (S,D,P) = \left(\begin{bmatrix} 1 & 0 \\ 1/\gamma & 1 \end{bmatrix}, \begin{bmatrix} \gamma & 1 \\ 0 & -1/\gamma \end{bmatrix}, I \right)$ and $\| D \|_1 = 1 + 1/\gamma$
- $D^{-1} = \begin{bmatrix} 1/\gamma & 1 \\ 0 & -\gamma \end{bmatrix}$ hence $\| D^{-1} \|_1 = \max\{ 1/\gamma, 1 + \gamma \}$ and $c_1(D) = (1 + 1/\gamma) \max\{ 1/\gamma, 1 + \gamma \}$

Then:

$$\lim_{\gamma \to 0} \frac{c_1(D)}{c_1(A)} = +\infty$$

So: by choosing a *sufficiently small* value of γ it is possible to obtain an amplification factor of the condition number *as large as desired*: the solution procedure of the system of equations which use EGP transforms the system $A\,x = b$ into the *equivalent* system $D\,x = c$ but while the conditioning properties of A are good ($c_1(A) < 4$) those of D, choosing γ suitably small, are *very bad* ($c_1(D)$ enormous).

While the solution procedure of the system of equations which uses EGP *is satisfactory* when operating in R (see (2.16) of Lecture 17), the procedure may be *unsatisfactory* when operating in F(β,m).

(2.48) Definition (EGPP procedure).

To overcome the potential danger highlighted in the previous example, a modification of the EGP procedure is used, leading to the definition of the EGPP procedure (Gaussian Elimination with Partial Pivoting). The difference with EGP lies *only* in *the choice of the permutation matrix* P_k. In the EGP procedure, we have:

if $A_k(k,k) \neq 0$ then $P_k = I$ otherwise
if there exists $i > k$ s.t. $A_k(i,k) \neq 0$ then $P_k = P_{k,i}$ otherwise $P_k = I$

In the EGPP procedure we proceed as follows:

if for every $i \geqslant k$ it is $A_k(i,k) = 0$ then $P_k = I$ otherwise
choose i s.t. $|A_k(i,k)| = \max \{ |A_k(j,k)| , j \geqslant k \}$ and set $P_k = P_{k,i}$

The choice in the EGP procedure is intended to ensure that the pivot is *non-zero.* In the EGPP procedure the aim is to have as pivot *the element of the* k*-th column of maximum modulus* among all those with row index $j \geqslant k$.

(2.49) Example.

Compute EGPP(A) with:

$$A = \begin{bmatrix} 1, & 0, & 1; \\ 2, & 1, & -1; \\ 1, & 2, & 1 \end{bmatrix}$$

(*) $A_1 = A$;

(*) $k = 1$; $|A_1(2,1)| = \max \{ |A_1(j,1)| , j \geqslant 1 \} \Rightarrow P_1 = P_{1,2}$;

$$T_1 = P_1 A_1 = \begin{bmatrix} 2 & 1 & -1 \\ 1 & 0 & 1 \\ 1 & 2 & 1 \end{bmatrix} , \quad H_1 = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_2 & 1 & 0 \\ \lambda_3 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{bmatrix}$$

The values λ_2,λ_3 are determined as in the EGP procedure.

Then:

$$H_1 T_1 = \begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & -1 \\ 1 & 0 & 1 \\ 1 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & -1 \\ 0 & -1/2 & 3/2 \\ 0 & 3/2 & 3/2 \end{bmatrix} = A_2$$

(*) $k = 2$; $|A_2(3,2)| = \max \{ |A_2(j,2)| , j \geqslant 2 \} \Rightarrow P_2 = P_{2,3}$;

$$T_2 = P_2 A_2 = \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3/2 & 3/2 \\ 0 & -1/2 & 3/2 \end{bmatrix} , \quad H_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/3 & 1 \end{bmatrix}$$

The value λ_3 is determined as in the EGP procedure.

Then:

$$H_2\,T_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/3 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3/2 & 3/2 \\ 0 & -1/2 & 3/2 \end{bmatrix} = \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3/2 & 3/2 \\ 0 & 0 & 2 \end{bmatrix} = A_3$$

(*) $D = A_3$; $P = P_2\,P_1$; $S = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & -1/3 & 1 \end{bmatrix}$ (determined as in EGP).

(2.50) Remark.

For any invertible matrix $A \in R^{n\times n}$, let (S,D,P) = EGPP(A). We have: $c_1(D)/c_1(A) \leqslant F(n)$. The function F depends *only* on the dimension n of the matrix and on the chosen norm, in particular it *does not depend on* A. Therefore, the growth factor of the condition number is *bounded*.

In the case of the Example (2.47) we have:

$$EGPP\left(\begin{bmatrix} \gamma & 1 \\ 1 & 0 \end{bmatrix}\right) = (S,D,P) \text{ with } D = I \Rightarrow c_1(D) = 1$$

(2.51) Remark (study using qr).

Let $A \in R^{n\times n}$ and $b \in R^n$. The procedure for studying the system A x = b using the qr procedure is:

(U,T) = qr(A);
if there exists k s.t. $t_{kk} = 0$ then STOP; otherwise
$c = U^t\,b$;
$x^* = SI(D,c)$
} in R

When using a computer, with set of machine numbers $F(\beta,m)$, the procedure transforms into:

$(\hat{U},\hat{T}) = qr_M(\hat{A})$;
if there exists k s.t. $\hat{t}_{kk} = 0$ then STOP; otherwise
$\hat{c} = \hat{U}^t \otimes \hat{b}$;
$\hat{x} = SI_M(\hat{T},\hat{c})$
} in $F(\beta,m)$

where $\hat{U}^t \otimes b$ is the column that is obtained by replacing in $U^t\,b$ the arithmetic operations with the corresponding predefined functions in $F(\beta,m)$.

(2.52) Example.

Similarly to what was done for the procedure using EGP, consider the following 'almost ideal' situation:

- $\hat{A} = A$, $\hat{b} = b$ - the components of the data are in $F(\beta,m)$;

- $qr_M(A) = qr(A) = (U,T)$ - the result of qr_M 'is exact', and T is invertible;
- $U^t \otimes b = \hat{c} = rd(c)$ - the result of $U^t \otimes b$ is 'almost ideal';
- $SI_M(T,\hat{c}) = SI(T,\hat{c})$ - the result of SI_M 'is exact'.

Under these hypotheses we have: $x^* = SI(T,c)$ is the solution of the system $T\,x = c$, $\hat{x} = SI(T,\hat{c})$ is the solution of the system $T\,x = \hat{c}$. Introducing the perturbation $\delta c = \hat{c} - c$ we have, using the two-norm (the 'natural' norm to use in R^n when using the QR factorization which brings into play the notion of orthogonality, therefore the scalar product in R^n, is the two-norm: the one induced by the scalar product):

$$N_2(\delta c) \leqslant u\, N_2(c) \qquad \text{hence} \qquad \varepsilon_c \leqslant u$$

By the Conditioning Theorem we still have:

$$\varepsilon_x \leqslant c_2(T)\, \varepsilon_c \leqslant c_2(T)\, u$$

and the upper limit of the relative measure of the deviation depends on the amplification factor of the condition number $c_2(T)/c_2(A)$.

But in this case we have:

- $A = U\,T \Rightarrow \| A \|_2 = \| U\,T \|_2 = \max \{ N_2(U\,T\,v),\ N_2(v) = 1 \} =^1 \max \{ N_2(T\,v),\ N_2(v) = 1 \} = \| T \|_2$
- $T^{-1} = A^{-1}\,U \Rightarrow \| T^{-1} \|_2 = \| A^{-1}\,U \|_2 = \max \{ N_2(\,A^{-1}\,U\,v),\ N_2(v) = 1 \} =^2 \max \{ N_2(A^{-1}\,w),\ N_2(U^t\,w) = 1 \} = \max \{ N_2(A^{-1}\,w),\ N_2(w) = 1 \} = \| A^{-1} \|_2$

Hence $c_2(T) = c_2(A)$, that is, the amplification factor of the conditioning number is now $c_2(T)/c_2(A) = 1$.

The solution procedure of the system of equations using qr is satisfactory *even* when operating in $F(\beta,m)$.

(2.3) COST OF THE SOLUTION'S COMPUTATION

(2.53) Definition (arithmetic cost).

One way to compare the two procedures described for obtaining an approximation of the solution of a system of linear equations (the one using EGPP and the one using qr) is to consider the time needed to compute the approximation.

In the context of solving systems of linear equations, we introduce the following notion of the *cost* of computing $\varphi(x)$, $C(\varphi)$, where φ is the *naive algorithm* (see Definition (1.32), Lecture 6) for f:

$$C(\varphi) = \text{the number of arithmetic operations needed to calculate f}$$

---

1 Since U is orthogonal we have: $N_2(U\,T\,v) = \text{sqrt}(\ v^t\,T^t\,U^t\,U\,T\,v\ ) = \text{sqrt}(\ v^t\,T^t\,T\,v\ ) = N_2(T\,v)$.

2 Change of variable: $w = U\,v$. Since U is orthogonal we have $v = U^t\,w$ e $N_2(U^t\,w) = N_2(w)$.

(2.54) Remark (the above one is a good definition).

For C(φ) to be representative of the *time* needed to compute φ(x), the following two conditions must be satisfied:

(1) When calculating φ(x), the time spent on activities *other than* the execution of arithmetic operations (i.e.: in the computation of predefined functions corresponding to elementary functions or comparisons) *must be negligible* (an example of an algorithm in which this condition is *not* verified is the one that calculates the infinite norm of a vector: in this case the algorithm only performs *comparisons* between the components of the vector);
(2) The computation time of each of the predefined functions corresponding to arithmetic operations must be *independent of the operands*.

The second condition is *not verified*, for example, in the case of the multiplication between two elements of F(β,m): to calculate $\xi_1 \otimes \xi_2$ it is necessary to *multiply the fractions* - and this occurs in a time independent of the factors because the fractions always have *the same number* of digits - and *add the exponents*; it is this last operation that cannot be considered independent of the factors because the exponents are just integer numbers that *have a number of digits that depends on which elements of* F(β,m) are considered. In particular, for the notion of arithmetic cost to be representative of the time required for the calculation, *the set of machine numbers of the computer must not be* F(β,m).