SISTEMI LINEARI: ANALISI DEGLI ERRORI

Gerardo Toraldo

Università di Napoli Federico II

A.A. 2014-2015

Forward Error Analysis

$$Ax = b, \quad x^* = A^{-1}b \tag{1}$$

Obiettivo della FEA: analizzare l'effetto di eventuali perturbazioni δA e δb su A e b rispettivamente, sulla soluzione calcolata di (1).

$$(A + \delta A)(x) = b + \delta b, \quad x^* + \delta x = (A + \delta A)^{-1}(b + \delta b); \tag{2}$$

Teorema (perturbazione solo su b)

.

$$\frac{1}{\kappa(A)}\frac{\|\delta b\|}{\|b\|} \leq \frac{\|\delta x\|}{\|x^*\|} \leq \kappa(A)\frac{\|\delta b\|}{\|b\|} \tag{3}$$

dimostrazione: da (2) segue che

$$\delta x = A^{-1}\delta b \tag{4}$$

Per dimostrare la seconda disequazione in (16) si osservi che

$$\|\delta x\| \leq \|A^{-1}\|\|\delta b\| = \|A^{-1}\|\|Ax\|\frac{\|\delta b\|}{\|b\|} \leq \|A^{-1}\|\|A\|\|x\|\frac{\|\delta b\|}{\|b\|}$$

Circa la prima disequazione, da (1) e (4) seguono le disequazioni:

$$\|\delta b\| \leq \|A\|\|\delta x\|; \quad \|x^*\| \leq \|A^{-1}\|\|b\|$$

Dividendole membro a membro

$$\frac{\|\delta b\|}{\|A^{-1}\|\|b\|} \leq \frac{\|A\|\|\delta x\|}{\|x^*\|}$$

$$A = \begin{pmatrix} 1001 & 1000 \\ 1000 & 1001 \end{pmatrix}, \; b = \begin{pmatrix} 1000 \\ 1000 \end{pmatrix}, \; b_1 = \begin{pmatrix} 1000 \\ 1001 \end{pmatrix} \; b_2 = \begin{pmatrix} 1000.71 \\ 1000.71 \end{pmatrix}$$

Si osservi che $\|b\| \approx 1414, \mathrm{cond}(A) \approx 2001, \delta b_1 \approx \delta b_2 \approx 1 \; \|\delta b_i\|/\|b\| \approx 7.07 \cdot 10^{-4}$ e quindi, in virtù di (16)

$$(3.533 \cdot 10^{-7}) \leq \frac{\|\delta x_i\|}{\|x^*\|} \leq (1.414) \; (i = 1, 2)$$

$$x^* = \begin{pmatrix} 0.499750124937489 \\ 0.499750124937574 \end{pmatrix}, \; x_1 = \begin{pmatrix} -0.000000000000057 \\ +1.000000000000057 \end{pmatrix}, \; x_2 = \begin{pmatrix} 0.500104947526235 \\ 0.500104947526239 \end{pmatrix}$$

$$\frac{\|\delta x_1\|}{\|x^*\|} = 1.00050, \; \frac{\|\delta x_2\|}{\|x^*\|} = 0.000071$$

Teorema (sistema perturbato su A e b)

Supponiamo in 2 sia $\delta A = 0$. Allora, per l'errore δx valgono le seguenti limitazioni:

$$\|A^{-1}\delta A\| < 1, \tag{5}$$

allora per l'errore δx vale la maggiorazione

$$\frac{\|\delta x\|}{\|x^*\|} \leq \frac{\kappa(A)}{1 - \|A^{-1}\|\|\delta A\|}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right) = \frac{\kappa(A)}{1 - \kappa(A)\|\delta A\|/\|A\|}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right) \tag{6}$$

$$(I + A^{-1}\delta A)(I + A^{-1}\delta A)^{-1} = I \Leftrightarrow (I + A^{-1}\delta A)^{-1} + A^{-1}\delta A(I + A^{-1}\delta A)^{-1} = I^*$$
$$\Leftrightarrow (I + A^{-1}\delta A)^{-1} = I - A^{-1}\delta A(I + A^{-1}\delta A)^{-1}$$

Passando alle norme nell'ultima uguaglianza si ha in particolare

$$\|(I + A^{-1}\delta A)^{-1}\| \leq 1 + \|A^{-1}\delta A\| \cdot \|(I + A^{-1}\delta A)^{-1}\| \Leftrightarrow \|(I + A^{-1}\delta A)^{-1}\|(1 - \|A^{-1}\delta A\|) \leq 1$$

e, se vale (5)

$$\|(I + A^{-1}\delta A)^{-1}\| \leq \frac{1}{1 - \|A^{-1}\delta A\|} \leq \frac{1}{1 - \|A^{-1}\|\|\delta A\|} \tag{7}$$

Da (2) si ha

$$Ax^* + A\delta x + \delta A x^* + \delta A\delta x = b + \delta b \Leftrightarrow (I + A^{-1}\delta A)\delta x = A^{-1}(\delta b - \delta A x^*)$$

*L'invertibilità della matrice $I + A^{-1}\delta A$ è conseguenza del teorema di Gastinel (si veda)

$$
\begin{aligned}
\delta x &= (I + A^{-1}\delta A)^{-1} A^{-1}(\delta b - \delta A x^*) \text{ e quindi, in virtú di (7)} \\
\|\delta x\| &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\delta A\|}(\|\delta b\| + \|\delta A\|\|x^*\|) \Leftrightarrow \\
\frac{\|\delta x\|}{\|x^*\|} &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\delta A\|}\left(\frac{\|\delta b\|}{\|x^*\|} + \|\delta A\|\right) \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\delta A\|}\left(\frac{\|\delta b\|}{\|b\|}\|A\| + \|\delta A\|\right) \text{ †}
\end{aligned}
$$

e quindi

$$
\frac{\|\delta x\|}{\|x^*\|} \leq \frac{\|A^{-1}\|\|A\|}{1 - \|A^{-1}\|\|\delta A\|}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right)
$$

da cui la tesi.

† $\|Ax^*\| = \|b\| \Rightarrow \|b\| \leq \|A\|\|x^*\| \Leftrightarrow \|x^*\| \geq \frac{\|b\|}{\|A\|}$

Un'osservazione sull'ipotesi 5. Si enuncia a tal proposito il seguente teorema

Teorema di Gastinel

Se A è una matrice non singolare, allora

$$\frac{1}{\kappa(A)} = \min_{\delta A} \left\{ \frac{\|\delta A\|}{\|A\|} \text{ tale che } A + \delta A \text{ è non singolare} \right\}$$

Da questo teorema, e da quelli precedentemente dimostrati segue che:

- se la matrice di perturbazione δA è tale che $\|\delta A\| < 1/\|A^{-1}\|$, allora le matrici $A + \delta A$ e $I + A^{-1}\delta A$ sono non singolari;
- una matrice malcondizionata è "quasi singolare"
- *piccole perturbazioni* (relative) sui dati producono *piccole perturbazioni* sulla soluzione. Per matrici malcondizionate le pertubazioni nei dati *possono essere fortemente amplificate nella soluzione*

$$A = \begin{pmatrix} 1001 & 1000 \\ 1000 & 1001 \end{pmatrix}, \quad \frac{1}{\kappa(A)} = 4.997501249375426 \cdot 10^{-4}$$

$$\delta A_1 = \begin{pmatrix} 0 & 0.99 \\ 0.99 & 0 \end{pmatrix}, \quad \frac{\|\delta A_1\|}{\|A\|} = 4.947526236881559 \cdot 10^{-4}$$

$$\delta A_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \frac{\|\delta A_2\|}{\|A\|} = 4.997501249375312 \cdot 10^{-4}$$

Errori di arrotondamento

Immaginando che i soli errori siano quelli dovuti all'utilizzo di un sistema a precisione finita,

$$\mathrm{fl}(b) = b + \delta b; \quad \mathrm{fl}(A) = A + \delta A,$$

è ragionevole assumere

$$\|\delta b\|_\infty \leq \mathbf{u}\|b\|_\infty, \ \|\delta A\|_\infty \leq \mathbf{u}\|A\|_\infty, \ ^{\ddagger}$$

Se $\mathbf{u}\kappa(A)_\infty \leq 1/2$, allora la (6) implica

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \leq 4\mathbf{u}\kappa(A)_\infty \ ^{\S}$$

$^{\ddagger}\mathbf{u} = 0.5\varepsilon$, dove $\varepsilon = \beta^{1-t}$ è l'unità di roundoff

§Golub, Van Loan *Matrix Computation* (3^{rd} edition), p.104

Errori di arrotondamento

Il seguente teorema fornisce una maggiorazione dell'errore sulla soluzione in ipotesi che chiaramente gneralizzano quelle considerate nel caso dell'errore di arrotondamento

Teorema[¶] Supponiamo che sia

$$\|\delta A\| < \gamma\|A\|,\ \|\delta b\| < \gamma\|b\| \quad \text{per un certo } \gamma > 0,\ \text{con } \gamma\kappa(A) < 1, \tag{8}$$

allora valgono le seguenti disuguaglianze:

$$\frac{\|x^* + \delta x\|}{\|x^*\|} \leq \frac{1+\gamma\kappa(A)}{1-\gamma\kappa(A)};\quad \frac{\|\delta x\|}{\|x^*\|} \leq \frac{2\gamma\kappa(A)}{1-\gamma\kappa(A)} \tag{9}$$

dimostrazione: Si osservi che dalle ipotesi segue che $(I + A^{-1})$ è invertibile.
Inoltre, da (2) segue che

$$\|x^* + \delta x\| \leq \|I + A^{-1}\delta A\|(\|A^{-1}b\| + \|A^{-1}\|\|\delta b\|) \tag{10}$$

Essendo

$$\|(I + A^{-1}\delta A)^{-1}\| \leq \frac{1}{1 - \|A^{-1}\delta A\|} \leq \frac{1}{1-\gamma\kappa(A)};\quad A^{-1}b = x^*;\quad \|\delta b\| \leq \gamma\|b\| \leq \gamma\|A\|\|x^*\|$$

da (10) segue la prima disequazione in (9). La seconda segue da (6)

[¶] Theorem 3.3 in ....

Errori di arrotondamento per una **matrice triangolare** T

Sia $T \in R^{n\times n}$ una matrice triangolare. Si puó dimostrare la seguente stima della matrice di perturbazione δT in funzione della sua dimensione n e delle caratteristiche del ***sistema floating point*** $\mathtt{F}(\beta, t)$ utilizzato (nell'ipotesi in cui sia se $0 < 1 - \mathbf{u} < 1$):

$$|\delta T| \leq \frac{n\mathbf{u}}{1 - n\mathbf{u}}|T| \quad \text{dove } \mathbf{u} = 0.5\beta^{1-t} \tag{11}$$

Si osservi che la disequazone (11) mette in relazione la *grandezza* degli elementi della matrice di pertubazione con *la grandezza* degli elementi di T, la dimensione della matrice e la precisione del sistema, ma non con il codizionamento di tale matrice. Utilizzando una espansione di Taylor in (11) si ottiene la disequazione

$$|\delta T| \leq n\mathbf{u}|T| \tag{12}$$

che, utilizzata nella seconda disequazione in (9)

$$\frac{\|\delta x\|}{\|x^*\|} \leq \frac{n\mathbf{u}\kappa(T)}{1 - n\mathbf{u}\kappa(T)} = n\mathbf{u}\kappa(T) + O(\mathbf{u}^2)$$

Analisi a posteriori

Sia y la soluzione calcolata del sistema; possiamo immaginare di scrivere

$$y = Cb,$$

con C approssimante di A^{-1}. L'analisi a posteriori si propone di stimare l'errore $e = y - x^*$ in funzione di y e C. Si definiscano:

$$R = AC - I, \quad r = b - Ay;$$

e sia $\|R\| < 1$. In tali ipotesi è possibile dimostrare che

$$\|A^{-1}\| \leq \frac{\|C\|}{1 - \|R\|} \tag{13}$$

Come conseguenza, essendo $e = -A^{-1}r$, si ha

$$\|e\| \leq \frac{\|r\|\|C\|}{1 - \|R\|} \tag{14}$$

Analisi a posteriori

In un contesto di analisi aposteriori, il bound

$$\frac{\|\delta x\|}{\|x^*\|} \leq \kappa(A) \frac{\|\delta b\|}{\|b\|} \tag{15}$$

può essere scritto come

$$\frac{\|e\|}{\|x^*\|} \leq \kappa(A) \frac{\|r\|}{\|b\|} \tag{16}$$

Metodo di Gauss

Circa l'accuratezza della soluzione calcolata con il **metodo di Gauss con pivoting** è possibile dimostrare che la soluzione calcolata y è soluzione esatta di un sistema perturbato

$$(A + \delta A)x = b \text{ con } \frac{\|\delta A\|}{\|A\|} \leq \rho n \varepsilon \tag{17}$$

dove il fattore di crescita $\rho = \max |U| / \max |A|$.
Quanto è grande il fattore di crescita? In teoria ρ potrebbe crescere come 2^{n-1}. In pratica è dell'ordine dell'unità.