Analisi degli Errori

Luca Gemignani
luca.gemignani@unipi.it

1 marzo 2018

Indice

Lezione 1: Errori nel Calcolo di una Funzione Razionale.

Sia $f\colon [a,b] \to R$ una funzione razionale e $x \in [a,b]$. Un algoritmo per il calcolo di $f(x)$ esprime tale valore come risultato di una sequenza di operazioni aritmetiche. Ad esempio

$$f(x) = \frac{x^2+1}{x} = ((x \cdot x) + 1)/x.$$

Errori nel calcolo di $f(x)$ vengono generati

1. dall'appprossimazione del dato x in macchina con la sua approssimazione $\tilde{x} \in \mathbb{F}(B,t,m,M)$;
2. dall'approssimazione della funzione f in macchina mediante una sua realizzazione g, ad esempio $g(x) = ((x \otimes x) \oplus 1) \oslash x$, espressa come corrispondente sequenza di operazioni aritmetiche di macchina.

La prima sorgente di errori conduce alla seguente definizione.

Definizione 1.1. Si dice *errore inerente* o *inevitabile* generato nel calcolo di $f(x) \neq 0$ la quantità

$$\epsilon_{in} = \frac{f(\tilde{x}) - f(x)}{f(x)}.$$

Si osserva che:

- L'errore inerente misura la sensibilità della funzione f e pertanto del *problema matematico* considerato rispetto alla perturbazione del dato iniziale. È indipendente dall'algoritmo (sequenza di operazioni aritmetiche) utilizzato per il calcolo di $f(x)$ e quindi per la risoluzione del problema matematico associato.

- Se l'errore inerente è qualitativamente elevato in valore assoluto diciamo che il relativo problema matematico è *mal condizionato*. Viceversa se l'errore inerente è qualitativamente modesto in valore assoluto diciamo che il relativo problema matematico è *ben condizionato*.

Esempio 1.1. Per il calcolo di $f(x) = \frac{x^2+1}{x}$ si ha

$$f(\tilde{x}) = \frac{\tilde{x}^2+1}{\tilde{x}} \doteq \frac{(x^2(1+2\epsilon_x)+1)(1-\epsilon_x)}{x} \doteq \frac{x^2+1}{x} + \epsilon_x(2x - \frac{x^2+1}{x}),$$

e quindi

$$|\epsilon_{in}| = |\frac{f(\tilde{x}) - f(x)}{f(x)}| \doteq |\frac{x^2-1}{x^2+1}||\epsilon_x| \leq u,$$

per cui il problema del calcolo di $f(x)$ risulta ben condizionato.

La seconda sorgente di errori conduce alla seguente definizione.

Definizione 1.2. Si dice *errore algoritmico* generato nel calcolo di $f(\tilde{x}) \neq 0$ la quantità

$$\epsilon_{alg} = \frac{g(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})}.$$

Si osserva che:

1. La funzione g dipende dall'algoritmo utilizzato per calcolare $f(x)$. Ad esempio per $f(x) = \frac{x^2+1}{x}$ potremmo avere $g(x) = g_1(x) = ((x \otimes x) \oplus 1) \oslash x$ come sopra oppure $g(x) = g_2(x) = x \oplus (1 \oslash x)$. In generale differenti algoritmi conducono a differenti errori algoritmici.

2. Se l'errore algoritmico è qualitativamente elevato in valore assoluto diciamo che l'algoritmo è numericamente *instabile*. Viceversa se l'errore algoritmico è qualitativamente modesto in valore assoluto diciamo che l'algoritmo è numericamente *stabile*.

Esempio 1.2. Per la valutazione dell'errore algoritmico nel calcolo di $f(x) = \frac{x^2+1}{x}$, $x \in \mathbb{F}(B, t, m, M)$, si considerano le implementazioni $g_1(x)$ e $g_2(x)$. Vale

$$g_1(x) \doteq (x^2(1+2\epsilon_1)+1)(1+\epsilon_2)(1+\epsilon_3)/x \doteq \frac{x^2+1}{x} + \frac{x^2+1}{x}(\epsilon_2+\epsilon_3) + 2x\epsilon_1,$$

da cui

$$|\epsilon_{alg_1}| = |\frac{g_1(x) - f(x)}{f(x)}| \doteq |(\epsilon_2+\epsilon_3) + 2\frac{x^2}{x^2+1}\epsilon_1| \leq 4u,$$

e pertanto il primo algoritmo risulta numericamente stabile. Riguardo il secondo algoritmo invece si ottiene:

$$g_2(x) = (x + \frac{1+\delta_1}{x})(1+\delta_2) \doteq x + 1/x + (x+1/x)\delta_2 + \delta_1/x,$$

da cui si ricava

$$|\epsilon_{alg_2}| = |\frac{g_2(x) - f(x)}{f(x)}| \doteq |\delta_2 + \delta_1/(x^2+1)| \leq 2u,$$

e pertanto si conclude che anche il secondo algoritmo è numericamente stabile. In altre situazioni la scelta dell'algoritmo di calcolo può risultare critica. Il lettore consideri ad esempio il caso in cui $f(x) = (x-1)/x = 1 - 1/x$.

Definizione 1.3. Si dice *errore totale* generato nel calcolo di $f(x) \neq 0$ mediante l'algoritmo specificato da g la quantità

$$\epsilon_{tot} = \frac{g(\tilde{x}) - f(x)}{f(x)}.$$

L'errore totale misura la differenza relativa tra l'output atteso e l'output effettivamente calcolato. In un'analisi al primo ordine vale

Teorema 1.1. Si ha $\epsilon_{tot} = \epsilon_{in} + \epsilon_{alg}$.

Dimostrazione. Vale

$$\epsilon_{tot} = \frac{g(\tilde{x}) - f(x)}{f(x)} = \frac{g(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})}\frac{f(\tilde{x})}{f(x)} + \frac{f(\tilde{x}) - f(x)}{f(x)} =$$
$$= \epsilon_{alg}(1 + \epsilon_{in}) + \epsilon_{in} \doteq \epsilon_{alg} + \epsilon_{in}.$$

□

Il teorema esprime il fatto che nel calcolo di una funzione razionale in un'analisi al primo ordine le due fonti di generazione degli errori individuate precedentemente forniscono contributi separati che possono essere analizzati indipendentemente. In un sistema floating point a precisione finita l'obiettivo dell'analisi numerica è pertanto quello di individuare algoritmi numericamente stabili per problemi ben condizionati.

Lezione 2: Tecniche per l'Analisi degli Errori.

La regolarità della funzione $f(x)$ ha implicazioni sulle proprietà del problema matematico da essa specificato. La continuità della funzione implica la *buona positura del problema*. Dalla relazione

$$\epsilon_{in} = \frac{f(\tilde{x}) - f(x)}{f(x)} = \frac{f(\tilde{x}) - f(x)}{\tilde{x} - x}\frac{1}{f(x)}\frac{\tilde{x} - x}{x},$$

si ricava che la differenziabilità di $f(x)$ è essenziale per il controllo dell'errore inerente. In particolare se assumiamo che $f(x)$ è derivabile due volte con continuità in (a, b) allora vale

$$f(\tilde{x}) = f(x) + f'(x)(\tilde{x} - x) + \frac{f''(\xi)}{2}(\tilde{x} - x)^2, \quad |\xi - x| \leq |\tilde{x} - x|,$$

da cui si ottiene

$$\epsilon_{in} = \frac{f(\tilde{x}) - f(x)}{f(x)} \doteq \frac{f'(x)}{f(x)} x\epsilon_x = c_x\epsilon_x, \quad c_x = \frac{f'(x)}{f(x)}x.$$

La quantità $c_x = \frac{f'(x)}{f(x)}x$ detta *coefficiente di amplificazione* fornisce una misura del condizionamento del problema. Più generalmente se $f\colon \Omega \to \mathbb{R}$ è definita su un insieme aperto di $\mathbb{R}^n$, differenziabile due volte su Ω ed il segmento di estremi $\tilde{\boldsymbol{x}}$ e $\boldsymbol{x}$ è contenuto in Ω allora vale

$$\epsilon_{in} = \frac{f(\tilde{\boldsymbol{x}}) - f(\boldsymbol{x})}{f(\boldsymbol{x})} \doteq \frac{1}{f(\boldsymbol{x})} \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\boldsymbol{x}) x_i \epsilon_{x_i}.$$

Esempio 2.1. Per $f(x) = (x^2 + 1)/x$ si ha

$$c_x = (2x - (x^2 + 1)/x) \cdot (x/(x^2 + 1)) \cdot x = (x^2 - 1)/(x^2 + 1).$$

Poichè $|c_x| \leq 1$ il problema del calcolo di $f(x)$ risulta ben condizionato.

Esempio 2.2. Per le operazioni aritmetiche si ottiene:

$$\begin{aligned}
&f(x, y) = x + y, \quad \epsilon_{in} \doteq c_x\epsilon_x + c_y\epsilon_y, \quad c_x = \frac{x}{x+y},\ c_y = \frac{y}{x+y};\\
&f(x, y) = x - y, \quad \epsilon_{in} \doteq c_x\epsilon_x + c_y\epsilon_y, \quad c_x = \frac{x}{x-y},\ c_y = -\frac{y}{x-y};\\
&f(x, y) = x \cdot y, \quad \epsilon_{in} \doteq c_x\epsilon_x + c_y\epsilon_y, \quad c_x = 1,\ c_y = 1;\\
&f(x, y) = x/y, \quad \epsilon_{in} \doteq c_x\epsilon_x + c_y\epsilon_y, \quad c_x = 1,\ c_y = -1.
\end{aligned}$$

Segue che la sottrazione di due numeri vicini tra loro è potenzialmente causa di elevata amplificazione degli errori relativi (di rappresentazione) cui sono soggetti gli addendi (*fenomeno della cancellazione numerica*). Differentemente le operazioni moltiplicative risultano ben condizionate. Ad esempio siano $x = 0.2178\ 10^2$ e $y = 0.218\ 10^2$ e si supponga di operarare con troncamento in base 10 con 3 cifre di rappresentazione ($u = 10^{-2}$). Si ha $\tilde{x} = 0.217\ 10^2$ e $\tilde{y} = y$. Pertanto $\tilde{x} \ominus \tilde{y} = -0.001\ 10^2 = -0.1$ mentre $x - y = -0.0002\ 10^2 = -0.2\ 10^{-1}$ e quindi $|\epsilon_{in}| = 0.8/0.2 = 0.4$.

L'analisi dell'errore algoritmico può basarsi sui risultati ottenuti per l'errore inerente. Si distinguono

1. tecniche di analisi *in avanti*;
2. tecniche di analisi *all'indietro*.

Per l'analisi in avanti dell'errore algoritmico si consideri uno step intermedio dell'algoritmo di calcolo ove abbiamo da calcolare il valore dell'operazione aritmetica $\psi(\tilde{x}, \tilde{y})$ a partire da due dati perturbati $\tilde{x} = x(1+\epsilon)$ e $\tilde{y} = y(1+\delta)$. Per quanto visto prima si ha

$$\psi(\tilde{x}, \tilde{y}) = \psi(x, y)(1 + c_x(\psi)\epsilon + + c_y(\psi)\delta + \gamma), \quad |\gamma| \leq u,$$

dove $c_x(\psi), c_y(\psi)$ e γ sono rispettivamente i coefficienti di amplificazione e l'errore locale dell'operazione ψ. Il calcolo dell'errore algoritmico totale può essere reso intuitivo con l'aiuto di un grafo ove i nodi corrispondono ai risultati intermedi generati dall'algoritmo. Se il risultato intermedio x corrisponde al nodo i mentre il risultato intermedio y corrisponde al nodo j allora l'operazione $\psi(x, y) = z$ genera un arco dal nodo i al nodo corrispondente a z con peso $c_x(\psi)$ ed un arco dal nodo j al nodo corrispondente a z con peso $c_y(\psi)$. Nel nodo corrispondente a z viene poi generato un nuovo errore locale γ riportato a fianco del nodo. L'errore algoritmico totale accumulato sul nodo corrispondente a z risulta dato dalla somma dell'errore locale più i pesi di arco moltiplicati per l'errore algoritmico totale accumulato sul nodo di origine dell'arco considerato in accordo alla relazione precedente. L'analisi in avanti dell'errore algoritmico conduce generalmente a valutazioni eccessivamente pessimistiche.

Per l'analisi all'indietro dell'errore algoritmico si assume che $g(\tilde{x}) \doteq f(\hat{x})$, ovvero che il valore effettivamente calcolato $g(\tilde{x})$ risulti uguale in un'analisi al primo ordine al valore assunto dalla funzione esatta f valutata in un dato perturbato $\hat{x}$. Se otteniamo una stima sull'errore $(\hat{x} - \tilde{x})/\tilde{x}$ (da qui l'appellativo *all'indietro*) allora siamo in grado di stimare l'errore algoritmico

$$\epsilon_{alg} = \frac{g(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})} = \frac{f(\hat{x}) - f(\tilde{x})}{f(\tilde{x})},$$

utilizzando i risultati per l'amplificazione dell'errore inerente. L'analisi all'indietro dell'errore algoritmico restituisce generalmente stime più realistiche ed eventualmente se possibile permette di concludere la stabilità dell'algoritmo in situazioni di buon condizionamento del problema. Trova ampia applicazione nell'analisi della stabilità degli algoritmi per l'algebra lineare numerica.

Esempio 2.3. Si consideri l'algoritmo $g(a, b) = (a \otimes a) \ominus (b \otimes b)$ per il calcolo di $f(a, b) = a^2 - b^2$. Si ha $g(a, b) \doteq a^2(1 + \epsilon_1 + \epsilon_3) - b^2(1 + \epsilon_2 + \epsilon_3)$ e quindi $g(a, b) = f(\hat{a}, \hat{b})$ con $\hat{a} = a\sqrt{1 + \epsilon_1 + \epsilon_3} \doteq a(1 + \delta_1)$ e $\hat{b} = b\sqrt{1 + \epsilon_2 + \epsilon_3} \doteq b(1 + \delta_2)$ con $|\delta_1| \leq u$ e $|\delta_2| \leq u$. Si confrontino dunque i risultati ottenuti per l'errore algoritmico con l'analisi in avanti (grafo) e l'analisi all'indietro.

Lezione 3: Cenni sul Calcolo di una Funzione non Razionale.

Nel calcolo di una funzione non razionale $h(x)$ si introduce un errore iniziale detto *errore di approssimazione* o *errore analitico* determinato dalla necessità di approssimare la funzione $h(x)$ con una funzione razionale $f(x)$.

Definizione 3.1. Si dice *errore analitico* generato nel calcolo di $h(x) \neq 0$ mediante la sua approssimazione razionale $f(x)$ la quantità

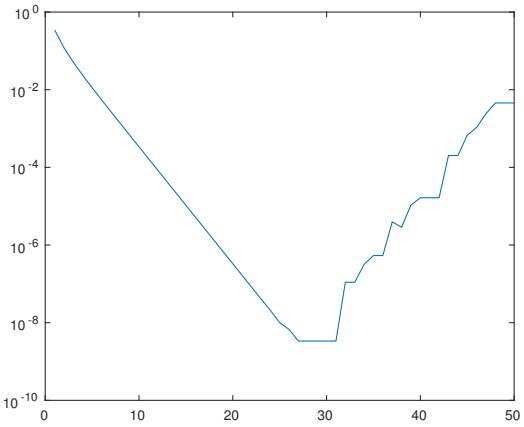
$$\epsilon_{an} = \frac{f(x) - h(x)}{h(x)}.$$

Nel calcolo della funzione $f(x)$ si genera poi un errore inerente ed algoritmico in accordo a quanto visto sopra per cui in un'analisi al primo ordine si perviene alla relazione

$$\epsilon_{tot} = \frac{g(\tilde{x}) - h(x)}{h(x)} \doteq \epsilon_{an} + (\epsilon_{in} + \epsilon_{alg}).$$

Il problema computazionale risulta pertanto quello di determinare approssimazioni razionali che consentano un bilanciamento tra le varie componenti presenti nella stima dell'errore totale.

Esempio 3.1. Si consideri il problema di approssimare la derivata della funzione $f(x) = \tan(x)$. Per x fissato si pone $g(h) = \frac{f(x+h)-f(x)}{h}$ e si considera l'approssimazione $g(h)$ di $f'(x)$ per valori di $h > 0$ decrescenti. Il seguente grafico riporta il plot dell'errore totale $\epsilon_{tot} = |\frac{\mathrm{fl}(g(h)) - f'(x)}{f'(x)}|$ per $x = 1/3$ e $h = 2^{-k}$, $1 \leq k \leq 50$. L'andamento del grafico rivela che per valori sufficiente-



mente piccoli di h l'errore inerente ed algoritmico commessi nel calcolo di $g(h)$ risultano predominanti rispetto alla riduzione dell'errore analitico. Per esercizio il lettore valuti l'errore inerente nel calcolo di $g(h)$ e l'errore algoritmico connesso nel calcolo di $g(h)$ assunto di disporre di valori calcolati in macchina per $f(x)$ e $f(x+h)$ con errore relativo limitato dalla precisione di macchina.