Cancellazione numerica e zeri di funzione

Dott. Marco Caliari

PLS a.s. 2012–2013

Capitolo 1

Aritmetica floating point

1.1 I numeri macchina

Data la capacità finita di un calcolatore, solo alcuni dei numeri reali possono essere rappresentati. I calcolatori usano la notazione *a virgola mobile (floating point)*: ogni numero $x = \pm 0.x_1x_2 \ldots x_tx_{t+1} \ldots \cdot 10^p$ (ove x_i è una cifra compresa tra 0 e 9) è rappresentato internamente come $\pm 0.x_1x_2 \ldots x_t \cdot 10^p$ ove p è una opportuna potenza (eventualmente negativa). Questa notazione sottintende che si sta usando la base 10 per la rappresentazione. In realtà, i calcolatori usano tipicamente la base 2. La capacità finita implica un limite sul numero, diciamo t, di cifre utilizzabili ("precisione") e sull'esponente p ("ordine di grandezza") che si assume variare tra $L < 0$ e $U > 0$. Inoltre, questa rappresentazione non è unica: per esempio, $0.1 \cdot 10^{-1} = 0.01 \cdot 10^0 = 0.001 \cdot 10^1$. Si assume quindi $x_1 \neq 0$ (rappresentazione *normalizzata*). L'insieme dei *numeri macchina* è allora

$$\{0\} \cup \{x \colon x = \text{sign}(x)(0.x_1x_2 \ldots x_t)_B \cdot B^p,\ 0 \leq x_i < B,\ x_1 \neq 0,\ L \leq p \leq U\}$$

ove B è la *base*, t il numero di *cifre significative*, $x_1x_2 \ldots x_t$ la *mantissa* e p la *caratteristica*. La scrittura $(0.x_1x_2 \ldots x_t)_B$ è una abbreviazione per

$$\sum_{i=1}^{t} x_i B^{-i} = \frac{x_1}{B^1} + \frac{x_2}{B^2} + \ldots + \frac{x_t}{B^t}$$

Nei calcoli si considera sempre una cifra in più, che determina poi l'*arrotondamento*, che avviene in maniera standard: si somma $B/2 \cdot B^{-(t+1)}$ al numero e si elimina la cifra x_{t+1}. In un normale calcolatore, $B = 2$ e $t = 52$. Viene considerata una cifra in più (per $t = 53$) solo per determinare l'arrotondamento. Dato un numero x e il suo arrotondamento a numero macchina $\text{fl}(x)$, si ha

$$|x - \text{fl}(x)| \leq \frac{B}{2} \cdot B^{-(t+1)} \cdot B^p = \frac{B}{2} \cdot B^{-t} \cdot B^{p-1} \leq \frac{B}{2} \cdot B^{-t} |x|$$

da cui

$$\frac{|x - \mathrm{fl}(x)|}{|x|} \leq \frac{B}{2} \cdot B^{-t} = \varepsilon$$

Il numero ε è chiamato *precisione di macchina* e vale $\varepsilon = 2^{-52} \approx 2.2204 \cdot 10^{-16}$. Il più grande numero rappresentabile come numero macchina è

$$(0.\underbrace{B-1B-1\ldots B-1B-1}_{t \text{ cifre}})_B \cdot B^U = (1 - (0.\underbrace{00\ldots01}_{t \text{ cifre}})_B) \cdot B^U = \\ = (1 - B^{-t}) \cdot B^U$$

In un normale calcolatore $U = 1023$ e dunque $(1 - B^{-t}) \cdot B^U \approx 8.9885 \cdot 10^{307}$. Il più piccolo numero rappresentabile come numero macchina è

$$(0.\underbrace{10\ldots00}_{t \text{ cifre}})_B \cdot B^L = B^{L-1}$$

In un normale calcolatore $L = -1022$ e dunque $B^{L-1} \approx 1.1125 \cdot 10^{-308}$.

Consideriamo ora una qualunque operazione aritmetica tra due numeri, per esempio la somma. La somma di due numeri x e y viene eseguita al calcolatore come

$$\mathrm{fl}(\mathrm{fl}(x) + \mathrm{fl}(y))$$

I due numeri vengono prima arrotondati in numeri macchina, ne viene eseguita la somma e, infine, il risultato è arrotondato a numero macchina. La stessa cosa succede per ogni altro tipo di operazione. Le due limitazioni dei numeri macchina, sulla precisione e sull'ordine di grandezza, possono portare a risultati sorprendenti, anche per operazioni apparentemente banali.

1.2 Cancellazione numerica

Consideriamo la seguente espressione

$$\frac{(1+x)-1}{x}$$

ove $x = 0.1234 \cdot 10^{-2}$. Il calcolo con numeri macchina con $t = 4$ (e $B = 10$) produce

$$(1+x) = 0.1001\boxed{2}\boxed{34} \cdot 10 \approx 0.1001\boxed{2} \cdot 10$$

che, arrotondato a t cifre decimali, è $0.1001 \cdot 10$. Poi

$$\boxed{(0.1001\text{-}0.1000)} \cdot 10 = 0.0001 \cdot 10 = 0.1000 \cdot 10^{-2}$$

E infine

$$(0.1000/0.1234) \cdot 10^{-2} = 0.8104$$

L'errore relativo commesso è $|1 - 0.8104| = 0.1896 \approx 20\%$. Il fatto che t sia comunemente molto più grande, non esclude l'insorgere di questo tipo di errori, detti di *cancellazione* (poiché alcune cifre sono state cancellate durante gli arrotondamenti). In particolare, è l'operazione sopra in riquadro ad essere incriminata, in quanto *sottrazione di due numeri vicini in modulo e approssimati*.

1.2.1 Radici di un'equazione di secondo grado

L'equazione $ax^2 + bx + c = 0$ ha una soluzione data da

$$x_1 = \frac{\sqrt{b^2 - 4ac} - b}{2a}$$

Presi $a = 10^{-10}$, $b = 1$, $c = 10^{-4}$, si ha $x_1 = -9.99200722162641 \cdot 10^{-5}$. Adesso calcoliamo $ax_1^2 + bx_1 + c$ ed otteniamo $7.99277837369190 \cdot 10^{-8}$ ben diverso dal risultato atteso. L'operazione incriminata qui è la sottrazione al numeratore tra due numeri approssimati e vicini in modulo. Se invece calcoliamo x_1 come

$$\hat{x}_1 = \frac{\sqrt{b^2 - 4ac} - b}{2a} \cdot \frac{\sqrt{b^2 - 4ac} + b}{\sqrt{b^2 - 4ac} + b} = \frac{-2c}{\sqrt{b^2 - 4ac} + b}$$

otteniamo $\hat{x}_1 = -1.00000000000001 \cdot 10^{-4}$ e $a\hat{x}_1^2 + b\hat{x}_1 + c$ produce 0.

1.2.2 Calcolo di π

Consideriamo un cerchio di raggio unitario. La sua area vale dunque π. Consideriamo ora un poligono regolare di 2^n lati inscritto nella circonferenza. Congiungendo i vertici del poligono con il centro della circonferenza, si ottengono 2^n triangoli equivalenti, isosceli, con i due lati uguali di lunghezza unitaria e con l'angolo al centro pari a $2\pi/2^n$ radianti. Dunque l'area di un singolo triangolo vale

$$T_{2^n} = \frac{1}{2} \sin \frac{2\pi}{2^n}$$

e dunque quella del poligono

$$P_{2^n} = 2^n \cdot \frac{1}{2} \sin \frac{2\pi}{2^n} = \frac{2^n}{2} \sqrt{1 - \cos^2 \frac{2\pi}{2^n}}$$

da cui

$$\cos\frac{2\pi}{2^n} = \sqrt{1 - \frac{4P_{2^n}^2}{2^{2n}}} = \sqrt{1 - 4^{1-n}P_{2^n}^2} \tag{1.1}$$

Il segno positivo davanti alla radice quadrata è corretto: infatti, ha senso parlare di poligono inscritto solo se $n \geq 2$, da cui $2\pi/2^n \leq \pi/2$ e dunque il suo coseno è non negativo. Consideriamo ora l'area del poligono regolare inscritto con $2 \cdot 2^n = 2^{n+1}$ lati: si ha

$$P_{2^{n+1}} = 2^{n+1}\frac{1}{2}\sin\frac{2\pi}{2^{n+1}} = 2^n \sin\frac{2\pi}{2 \cdot 2^n}$$

Usando la formula di bisezione del seno, si ha

$$P_{2^{n+1}} = 2^n\sqrt{\frac{1 - \cos\frac{2\pi}{2^n}}{2}}$$

e inserendo la (1.1), si trova la relazione ricorsiva

$$P_{2^{n+1}} = 2^n\sqrt{\frac{1 - \sqrt{1 - 4^{1-n}P_{2^n}^2}}{2}}$$

che si può scrivere

$$P_{2^{n+1}} = 2^{n-1/2}\sqrt{1 - \sqrt{1 - 4^{1-n}P_{2^n}^2}}, \quad n \geq 2$$

Dunque, l'area del poligono di 2^{n+1} lati si può ricavare dall'area del poligono di 2^n lati con semplici operazioni aritmentiche e estrazioni di radici quadrate. In particolare, per $n = 2$, si ha l'area del quadrato inscritto che vale 2. Per n grande, l'area del poligono di 2^{n+1} lati tende all'area del cerchio, dunque a π. Dunque, cambiando i nomi,

$$\begin{aligned} P_2 &= z_1 = 0 \\ P_4 &= z_2 = 2 \\ P_{2^{n+1}} &= z_{n+1} = 2^{n-1/2}\sqrt{1 - \sqrt{1 - 4^{1-n}z_n^2}}, \quad n \geq 2 \end{aligned}$$

si ha che z_{n+1} tende a π per n grande. Cosa succede se tentiamo di calcolare z_{n+1} per n grande? Si hanno errori di *cancellazione numerica*.

Per risolvere il problema nel calcolo di π, occorre razionalizzare, cioè calcolare z_{n+1} come

$$z_{n+1} = z_{n+1} \cdot \frac{\sqrt{1+\sqrt{1-4^{1-n}z_n^2}}}{\sqrt{1+\sqrt{1-4^{1-n}z_n^2}}} = \frac{2^{n-1/2}\sqrt{1-(1-4^{1-n}z_n^2)}}{\sqrt{1+\sqrt{1-4^{1-n}z_n^2}}} =$$
$$= \frac{2^{n-1/2}\sqrt{4^{1-n}z_n^2}}{\sqrt{1+\sqrt{1-4^{1-n}z_n^2}}} = \frac{\sqrt{2}z_n}{\sqrt{1+\sqrt{1-4^{1-n}z_n^2}}}$$

1.3 Overflow e underflow

Operando con numeri molto grandi, vicini al massimo numero rappresentabile $(1 - B^{-t}) \cdot B^U$, si possono generare numeri maggiori del massimo numero rappresentabile e appartenenti alla cosiddetta regione di *overflow*. Se per esempio a e b sono molto vicini al massimo numero rappresentabile, non è possibile calcolare esplicitamente

$$\frac{a+b}{2}$$

ma bisogna per forza fare

$$\frac{a}{2} + \frac{b}{2}$$

Analogamente per il calcolo

$$\sqrt{a^2+b^2}$$

che può essere ricondotto a

$$\max\{a,b\} \cdot \sqrt{\left(\frac{a}{\max\{a,b\}}\right)^2 + \left(\frac{b}{\max\{a,b\}}\right)^2}$$

Analogamente, l'insieme dei numeri diversi da zero e minori del più piccolo numero macchina (o maggiori del suo opposto) si chiama regione di *underflow*. Un risultato in regione di underflow viene comunemente arrotondato a zero.

Capitolo 2

Zeri di funzione

Consideriamo tre metodi per la ricerca di zeri di funzioni $f(x)$ in un intervallo $[a, b]$. Per semplicità, consideriamo che la funzione f abbia un solo zero nell'intervallo.

2.1 Metodo di bisezione

È molto semplice e richiede solo che f sia continua nell'intervallo $[a, b]$ e che cambi segno agli estremi. Si definisce

$$x_1 = \frac{a+b}{2}$$

e si valuta $f(x_1)$. Se $f(a)f(x_1) < 0$ significa che la funzione cambia segno nell'intervallo $[a, x_1]$ e si pone $b = x_1$. Se invece $f(a)f(x_1) > 0$ significa che la funzione cambia segno nell'intervallo $[x_1, b]$ e si pone $a = x_1$. A questo punto

$$x_2 = \frac{a+b}{2}$$

e si itera il processo. All'iterazione k l'intervallo di ricerca dello zero risulta avere ampiezza $(b-a)/2^k$. Pertanto, detto ξ lo zero di f, l'errore all'iterazione k soddisfa

$$|e_k| = |x_k - \xi| \leq \frac{b-a}{2^k}$$

e possiamo arrestare il calcolo quando questa stima è minore di una tolleranza prefissata. Notiamo anche che la stima equivale allo *scarto*

$$|x_k - x_{k-1}| = \frac{b-a}{2^k}, \quad k \geq 2$$

Ad ogni modo, la stima d'errore del metodo di bisezione potrebbe essere una sovrastima dell'errore vero. Possiamo dire che

$$\frac{|e_{k+1}|}{|e_k|} \simeq \frac{1}{2}$$

cioè la convergenza è *lineare*. Se ad un certo punto

$$\frac{|x_k - \xi|}{|\xi|} \approx 0.1$$

(una cifra significativa corretta), allora servono altre $\log_2 10$ iterazioni per avere 2 cifre corrette, altre $3 \cdot \log_2 10$ per averne quattro corrette e così via.

Sembrerebbe cosa buona usare come criterio d'arresto anche il seguente: se $f(x_k)$ è minore della tolleranza prefissata, allora fermiamoci. Ma consideriamo il teorema del valor medio di Lagrange

$$f(x_k) - f(\xi) = f'(\xi_k)(x_k - \xi), \quad \xi_k \text{ "tra" } \xi \text{ e } x_k$$

da cui

$$|x_k - \xi| = \frac{|f(x_k)|}{|f'(\xi_k)|}$$

Dunque l'errore è proporzionale non al *residuo* $f(x_k)$ ma piuttosto al *residuo pesato* $f(x_k)/f'(\xi_k)$. Ovviamente non si conosce ξ_k e dunque come criterio di arresto si può valutare

$$\frac{|f(x_k)|}{|f'(x_k)|}$$

e arrestare il calcolo quando minore della tolleranza prefissata.

2.2 Metodo di Newton

Dato un punto x_k, sostituiamo $f(x)$ con la retta tangente a f in x_k

$$f(x_k) + f'(x_k)(x - x_k)$$

e definiamo x_{k+1} come lo zero di questa nuova funzione

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

Abbiamo il metodo di Newton. Ovviamente è necessario che la funzione f sia derivabile. In realtà serve qualcosa in più. Proviamo a studiare l'errore, sviluppando $f(\xi)$ is serie di Taylor con resto di Lagrange

$$f(\xi) = f(x_k) + f'(x_k)(\xi - x_k) + \frac{f''(\xi_k)}{2}(\xi - x_k)^2$$

da cui

$$-f(x_k) + (x_k - \xi)f'(x_k) = \frac{f''(\xi_k)}{2}(x_k - \xi)^2$$

e dunque

$$-\frac{f(x_k)}{f'(x_k)} + x_k - \xi = \frac{f''(\xi_k)}{2f'(x_k)}(x_k - \xi)^2$$

cioè

$$e_{k+1} = \frac{f''(\xi_k)}{2f'(x_k)}e_k^2$$

Supponiamo $|f''(x)/(2f'(x))| \leq M$ in un intorno sufficientemente grande di ξ: allora

$$\begin{aligned} e_2 &\leq Me_1^2 = \frac{1}{M}(Me_1)^2 \\ e_3 &\leq Me_2^2 = \frac{1}{M}(Me_2)^2 \leq \frac{1}{M}(Me_1)^{2^2} \\ &\dots \\ e_k &\leq \frac{1}{M}(Me_1)^{2^{k-1}} \end{aligned}$$

Quindi il metodo converge se $Me_1 < 1$, cioè se il punto di partenza x_1 è sufficientemente vicino a ξ. Sembra una contraddizione, ma nella realtà ci sono situazioni anche più favorevoli. Siccome

$$\frac{e_{k+1}}{e_k^2} \simeq M$$

la convergenza del metodo è *quadratica*. Ciò significa che se ad un certo punto

$$\frac{|x_k - \xi|}{|\xi|} \approx 0.1$$

con altre quattro iterazioni l'errore relativo è alla precisione di macchina.

Come criterio d'arresto si usa lo *scarto* $x_{k+1} - x_k$: infatti

$$f(x_k) - f(\xi) = f'(\xi_k)(x_k - \xi)$$

da cui

$$|x_{k+1} - x_k| = \left|\frac{f(x_k)}{f'(x_k)}\right| \approx \left|\frac{f(x_k)}{f'(\xi_k)}\right| = |e_k|$$

2.2.1 Estrazione di radici quadrate

Vogliamo calcolare la radice di $a > 0$, $a \neq 1$. Basta costruire un quadrato di area a e prenderne il suo lato. Partiamo da un rettangolo di base 1 e altezza a. Non assomiglia per niente al quadrato che vogliamo costruire: proviamo a prendere un rettangolo di base $(1 + a)/2$ (la media tra i lati del rettangolo precedente) e altezza in modo che l'area sia a. Se il risultato ottenuto è abbastanza quadrato, la base sarà una buona approssimazione di $\sqrt{a}$, altrimenti ripetiamo. Dette b_k e h_k la base e l'altezza all'iterazione k ($b_k h_k = a$), abbiamo

$$b_{k+1} = \frac{b_k + h_k}{2}$$
$$h_{k+1} = \frac{a}{b_{k+1}}$$

da cui

$$b_{k+1} = \frac{b_k + \frac{a}{b_k}}{2} = \frac{b_k}{2} + \frac{a}{2b_k}$$

Vediamo il metodo di Newton applicato al calcolo di radici quadrate. Consideriamo la funzione $f(x) = x^2 - a$, $a > 0$. Calcolarne lo zero significa calcolare la radice quadrata di a. Si pone $x_1 = a$ se $a > 1$, altrimenti $x_1 = 1/a$. Poi

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^2 - a}{2x_k} = \frac{2x_k^2 - x_k^2 + a}{2x_k} = \frac{x_k}{2} + \frac{a}{2x_k}$$

e si ritrova il metodo del rettangolo.

2.3 Metodo delle secanti

Una modifica del metodo di Newton è la seguente: dati due punti iniziali x_{k-1} e x_k si sostituisce $f(x)$ con la retta secante

$$f(x_k) + \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}(x - x_k)$$

e si definisce x_{k+1} come lo zero di questa nuova funzione

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}$$

Le ipotesi di convergenza sono quelle del metodo di Newton, ma non serve il calcolo esplitico della derivata. Si dimostra che l'ordine di convergenza è $(1 + \sqrt{5})/2$.