Appunti di Analisi Numerica*

Giuseppe Profiti

19 ottobre 2006

1 Arrotondamento in standard IEEE

L'arrotondamento è definito come arrotondamento ai pari. Dati $X \leq \alpha < Y$

$$fl_A(\alpha) = \begin{cases} X & \text{se } \alpha < \frac{X+Y}{2} \\ pari(X,Y) & \text{se } \alpha = \frac{X+Y}{2} \\ Y & \text{se } \alpha < \frac{X+Y}{2} \end{cases}$$

In base 2 l'ultima cifra è quindi 0 (pari).

Questo arrotondamento ha effetto sulla caratterizzazione di u che diventa il più grande v che sommato a 1 resta 1.

$$u = max(v | fl_A(v+1) = 1)$$

2 Algoritmo per trovare u

Arrotondamento con programma in C

```
int t;
float eps,somma;
eps = 1.0;
t = 0;
somma = 2.0;
while (somma > 1.0) {
eps = 0.5 * eps;
```

```
somma = eps + 1.0;
t++; }
```

Con *float* si ha la versione single, con *double* quella double.

- *eps* memorizza le potenze negative di 2
- t memorizza il numero di cifre

Mettendo tutto nel while il risultato cambia in base al'implementazione. Se si usano meno operandi rispetto al numero di registri i valori non vengono portati in memoria e t indicherebbe il numero di cifre dei registri.

Valori di u

- basic single $u = 2^{-23} \approx 10^{-7}$
- basic double $u = 2^{-53} \approx 10^{-16}$

Risultati del programma[1]:

- basic single t = 24 e ϵ = 5.960464e-08
- basic double t = 53 e ϵ = 1.110223e-16

Dato che lavoriamo in decimale vogliamo sapere quanti decimali corrispondono alla nostra precisione in binario.

$$\begin{aligned} (t)_2 &\Rightarrow (s)_1 0 \\ 2^{-t} &= 10^{-s} \\ log_{10} 2^{-t} &= log_{10} 10^{-s} \\ s &= t \cdot log_1 02 = t \cdot 0.30103 \end{aligned}$$

- basic single t = 24 e s = 7.224...
- basic double t = 53 e s = 15.954...

[1] il link è vicino a quello di questo documento

3 Aritmetica floating point

3.1 Precisione

Ogni operazione è implementata in modo che il risultato sia esatto a meno della conversione in un numero finito.

$a, b \in \mathcal{F}$ e $\bigcirc$ operazione in aritmetica macchina, *op* operazione esatta.

$$a \bigcirc b = fl(a \text{ op } b)$$

Da questo possiamo definire l'errore di calcolo

$$a \bigcirc b = (a \text{ op } b)(1 + \epsilon) \text{ con } |\epsilon| < u$$

$$\left| \frac{fl(a \text{ op } b) - (a \text{ op } b)}{(a \text{ op } b)} \right| < u$$

u è sia la precisione di rappresentazione sia la precisione di calcolo.

3.2 Operazioni

In aritmetica finita non valgono alcune proprietà, ad esempio quella associativa.

Non valgono:

- associativa di somma e moltiplicazione
- semplificazione $\frac{a \cdot b}{a} \neq b$ in generale
- cancellazione $a \cdot b = b \cdot c \not\Rightarrow a = c$

I confronti tra a e b non hanno senso numericamente, $a = b$ a meno dell'errore di calcolo

$$abs(a - b) \leq u$$

3.2.1 Esercizio

In $\mathcal{F}(10, 2, \lambda, \omega)$ verificare $(a \oplus b) \oplus c \neq a \oplus (b \oplus c)$.

Con a = 0.11, b = 0.013, c = 0.014.

$$\begin{aligned} (0.11 \oplus 0.013) \oplus 0.014 &\neq 0.11 \oplus (0.013 \oplus 0.014) \\ 0.12 \oplus 0.014 &\neq 0.11 \oplus (0.027) \\ 0.13 &\neq 0.14 \end{aligned}$$

4 Analisi degli errori

L'analisi e la stima degli errori è sempre nel caso peggiore.

- Analisi in avanti: si fa la stima a ogni operazione che si compie. É buona ma ha problemi con calcoli complicati

$$x \overset{f}{\longmapsto} f(x) \qquad x \overset{alg}{\longmapsto} alg(x) \qquad \text{differenza tra } f(x) \text{ e } alg(x)$$

- :Analisi all'indietro: dal risultato del calcolo trovo i dati possibili di origine, è più semplice perché non tiene conto di tutte le stime delle singole operazioni.

$$x \overset{f}{\longmapsto} f(x) \qquad f(x) \to x + \delta x$$

Esempio in avanti

$$\left| \frac{fl(a+b) - (a+b)}{a+b} \right| < u$$

Su 1 operazione la stima è u.

Esempio all'indietro

$$f(a+b) = (a+b)(1+\epsilon) = a(1+\epsilon) + b(1+\epsilon)$$

$|\epsilon| < u$, il risultato è accettabile.

4.0.2 Esempio 1: moltiplicazione tra reali

Analisi in avanti sapendo che:

$$\begin{aligned} fl(x) &= x \cdot (1+\epsilon_1) \\ fl(y) &= y \cdot (1+\epsilon_2) \\ |\epsilon_1|, |\epsilon_2|, |\epsilon_3| &< u \\ fl(fl(x) \cdot fl(y)) &= fl(x) \cdot fl(y) \cdot (1+\epsilon_3) \end{aligned}$$

$$\begin{aligned} \frac{fl(fl(x) \cdot fl(y)) - x \cdot y}{x \cdot y} &= && (1) \\ &= \frac{(x(1+\epsilon_1) \cdot y(1+\epsilon_2))(+\epsilon_3) - x \cdot y}{x \cdot y} && (2) \\ &= (1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3) - 1 && (3) \\ &= \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_1\epsilon_2 + \epsilon_1\epsilon_3 + \epsilon_2\epsilon_3 + \epsilon_1\epsilon_2\epsilon_3 && (4) \\ &\cong \epsilon_1 + \epsilon_2 + \epsilon_3 < 3u && (5) \end{aligned}$$

La moltiplicazione è un'operazione tranquilla.
Note:

- prima si approssimano i valori e poi si approssima il risultato
- si semplifica $x \cdot y$ trai passaggi 2 e 3
- si fa una stima del primo ordine trai passaggi 4 e 5 ($u^N < u$)

4.0.3 Esempio 2: addizione tra reali

$$\begin{aligned}
\frac{fl(fl(x)+fl(y))-(x+y)}{x+y} &= \\
&= \frac{(x(1+\epsilon_1)+y(1+\epsilon_2))(1+\epsilon_3)-(x+y)}{x+y} \\
&= \frac{x+y+x\epsilon_1+y\epsilon_2+x\epsilon_3+y\epsilon_3+x\epsilon_1\epsilon_3+y\epsilon_2\epsilon_3-(x+y)}{x+y} \\
&\cong \frac{x}{x+y}\epsilon_1+\frac{y}{x+y}\epsilon_2+\epsilon_3
\end{aligned}$$

Se x e y sono opposti in segno e di valori molto vicini ho che i 2 coef. delle ϵ sono molto grandi.

I.e. se $\frac{x}{x+y} = 10^7$ ho un errore dell'ordine di $10^7 \cdot u$.

L'operazione è critica, è un caso di cancellazione numerica.

Esempio numerico

$\mathcal{F}(10, 6, \lambda, \omega)$
$x = 0.147554326$
$y = -0.147251742$
$x + y = 0.302584 \times 10^{-3}$

$fl(x) = 0.147554$
$fl(y) = -0.147252$

$fl(fl(x) + fl(y)) = 0.000302 = 0.302 \times 10^{-3}$

$\left|\frac{0.302\times10^{-3}-0.302584\times10^{-3}}{0.302584\times10^{-3}}\right| = 0.00193 \cong 0.2 \times 10^{-2}$

Si confronta con $u = 0.5 \times 10^{-}5$, l'errore è molto grande (3 ordini di grandezza).

il problema nasce dall'arrotondamento di x e y, facendo la somma ottengo degli 0, quindi ho un risultato di sole 3 cifre, mentre le successive potrebbero servirmi. La cancellazione è quindi la sparizione di queste cifre utili (i.e. dalla quarta in poi).

5 Condizionamento di un problema e stabilità di un algoritmo

Modelliamo un problema come $f : \Re \to \Re$ (successivamente vedremo $f : \Re^n \to \Re$ e $f : \Re^n \to \Re^m$).

Abbiamo che, posto $\tilde{x} = fl(x)$

$$x \to f(x) \tag{6}$$
$$\tilde{x} \to f(\tilde{x}) \tag{7}$$
$$\tilde{x} \to \psi(\tilde{x}) \tag{8}$$

6 è la funzione reale, 7 è il risultato di f su dati finiti, 8 è il risultato in aritmentica finita. L'ideale sarebbe che $f(\tilde{x}) = \psi(\tilde{x})$.

Definizione 1 (di errore inerente) *L'errore inerente è quello che si ha utilizzando i numeri finiti. Non è eliminabile.*

$$E_{IN} = \left| \frac{f(\tilde{x}) - f(x)}{f(x)} \right|$$

Definizione 2 (di errore algoritmico) *L'errore algoritmico è quello dovuto alle operazioni rispetto al risultato con i numeri finiti.*

$$E_{ALG} = \left| \frac{\psi(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})} \right|$$

Definizione 3 (di errore totale) *L'errore totale indica quanto ci si è discostati dalla soluzione ideale.*

$$E_{TOT} = \left| \frac{\psi(\tilde{x}) - f(x)}{f(x)} \right|$$

Se l'errore algoritmico è contenuto allora l'algoritmo è stabile.

Se l'errore inerente è contenuto allora il problema è ben condizionato.

Teorema 1 *Dato $f : \Re \to \Re$ e $x, \tilde{x}$ t.c. $f(x), f(\tilde{x}) \neq 0$ allora:*

$$E_{TOT} = E_{ALG}(1 + E_{IN}) + E_{IN} \cong E_{ALG} + E_{IN}$$

Se E_{IN} è grande non ha senso migliorare E_{ALG}. Se E_{IN} è piccolo devo trovare un algoritmo tale che E_{ALG} sia piccolo.

Definizione 4 (di numero di condizione) *Dato ϵ_d errore sui dati*

$$NC = \frac{E_{IN}}{\epsilon_d}$$

Se NC è dell'ordine di 1, è ben condizionato, se è molto grande è mal condizionato.

$$\epsilon_d = \left|\frac{\tilde{x} - x}{x}\right| = \left|\frac{x + h - x}{x}\right| = \left|\frac{h}{x}\right|$$

$$E_{IN} = \left|\frac{f(x+h) - f(x)}{f(x)}\right|$$

$$NC = \left|\frac{f(x+h) - f(x)}{f(x)}\right| \cdot \left|\frac{x}{h}\right| = \left|\frac{f(x+h) - f(x)}{h}\right| \cdot \left|\frac{x}{f(x)}\right|$$

$$\lim_{h \to 0} \frac{E_{IN}}{\epsilon_d} = \frac{x f'(x)}{f(x)}$$