Computazione I

Paolo Bettelini

Contents

1 Floating points

L'insieme dei floating point è

$$f(\beta,t,m,M)=\{0,\text{NaN},\pm\infty\}\cup\left\{x=\text{sign}(x)\cdot\beta^e\sum_{i=1}^{t}y_i\beta^{-i}\;\middle|\;t,y_i,m,M\in\mathbb{N},y_1\neq 0,-m\leq e\leq M\right\}$$

Stimiamo ora l'errore relativo

$$\frac{|x-\tilde{x}|}{|x|}$$

dove $x\in\mathbb{R}$ e $\tilde{x}\in f(\beta,t,m,M)$ è la sua rappresentazione migliore in un calcolatore. Consideriamo $x>0$. Chiaramente, se $\tilde{x}\in\mathbb{R}$, allora $|x-\tilde{x}|=0$. Altrimenti, $x\in[a,b]$ dove $a,b\in f$ e sono consecutivi in f. Quindi

$$|x-\tilde{x}|\leq\frac{b-a}{2}$$

Abbiamo allora

$$a=\beta^e\sum_{i=1}^{t}y_i\beta^{-i}$$

e

$$b=\beta^e\left(\sum_{i=1}^{t}y_i\beta^{-i}+\beta^{-t}\right)=a+\beta^{e-t}$$

Quindi la differenza è data da

$$|x-\tilde{x}|\leq\frac{1}{2}\beta^{e-t}$$

Dobbiamo ora minorare l'elemento normalizzante

$$|x|=\beta^e\sum_{i=1}^{\infty}y_i\beta^{-i}\geq\beta^e\cdot y_1\beta^{-1}\geq\beta^{e-1}$$

Abbiamo quindi

$$\frac{1}{|x|}\leq\beta^{1-e}$$

Combinando i due risultati otteniamo

$$\frac{|x-\tilde{x}|}{|x|}\leq\frac{1}{2}\beta^{e-t}\beta^{1-e}=\frac{1}{2}\beta^{1-t}\triangleq u$$

Allora u è la precisione macchina.

2 Approssimazione di zeri di funzioni

Sia $f \in C_{[a,b]}$ tale che $f(\alpha) = 0$. Vogliamo approssimare α numericamente.

3 Metodo di bisezione

Possiamo applicare ricorsivamente il teorema degli zeri, quindi bisezione. In questo caso la velocità è indipendente da f ma solo dipendente dalla grandezza dell'intervallo. Terminiamo l'algoritmo quando $|b-a| < \varepsilon$ che è la mia tolleranza. L'errore relativo è $|x-\alpha| < \varepsilon|\alpha|$. Per trovare il numero di operazioni abbiamo

$$|b_1 - a_1| = \frac{1}{2}|b_2 - a_1|, \cdots, |b_i - a_i| = \frac{1}{2^i}|b_i - a_i|$$

Quindi servono

$$\left\lceil \log 2 \left(\frac{|b-a|}{\varepsilon} \right) \right\rceil$$

Il pro di questo metodo è quindi una convergenza globale ma come contro abbiamo una convergenza lenta se l'intervallo è grande.

4 Metodo di iterazione funzionale

Si definiscono metodi numerici per generare la successione $\{x_k\}$ tale che possibilmente

$$\lim_k x_k = \alpha$$

Si andranno a definire iterazioni funzionali della forma

$$x_{k+1} = g(x_k)$$

con un x_0 dato. Vogliamo convertire $f(x) = 0$ in un'equazione di punto fisso $x = g(x)$, e poi si definisce l'iterazione. L'iterazione funzionale deve tuttavia convergere. Quindi, data g sufficientemente regolare tale che $\alpha = g(\alpha)$ e definito lo schema di iterazione $x_{k+1} = g(x_k)$ con x_0 dato, si vogliono definire condizioni necessarie e/o sufficienti per la convergenza

$$\lim_k x_k = \alpha$$

La condizione necessaria è

Lemma

Sia $g \in C([a,b])$ tale che $x_0 \in [a,b]$ e $x_{k+1} = g(x_k) \in [a,b]$ e

$$\lim_k x_k = \alpha$$

allora $\alpha = g(\alpha)$.

Proof

$$\alpha = \lim_k x_k = \lim_k x_{k+1} = \lim_k g(x_k)$$

La condizione sufficiente è il teorema delle contrazioni, per la quale serve $f \in C^1$.

Teorema Teorema delle contrazioni semplificato

Sia $g \in C^1(I_\delta(\alpha))$ dove $g(\alpha) = \alpha$. Sia $x_0 \in I_\delta(\alpha)$ e

$$|g'(x)| < 1, \quad \forall x \in I_\delta(\alpha)$$

Allora per la $\{x_k\}$ tale che $x_{k+1} = g(x_k)$ vale

1. $x_k \in I_\delta(\alpha)$;
2. $$\lim_k x_k = \alpha$$
3. α è l'unico punto fisso.

Proof Teorema delle contrazioni semplificato

1. Per induzione su k
   - il caso base $k = 0$ è banale per ipotesi;
   - assumendo che x_k dimostriamo che x_{k+1} sta nell'intorno.

   $$x_{k+1} \in I_\delta(\alpha) \iff |x_{k+1} - \alpha| \leq \delta$$

   Per il teorema di Lagrange sul punto ξ_k

   $$\begin{aligned} |x_{k+1} - \alpha| = |g(x_k) - g(\alpha)| &= |g'(\xi_k)(x_k - \alpha)| \\ &= \underbrace{|g'(\xi_k)|}_{<1} \cdot \underbrace{|x_k - \alpha|}_{\leq \delta} \\ &< \delta \end{aligned}$$

   con $|\xi_k - \alpha| < |x_k - \alpha|$. Quindi appartiene all'intervallo.
2. Sia
   $$\lambda = \max_{x \in I_\delta(\alpha)} |g'(x)| < 1$$
   Sia anche $e_k = |x_k - \alpha|$ l'errore. Siccome $e_{k+1} \leq \lambda e_k$ abbiamo
   $$\lim_k x_k = \alpha \iff \lim_k e_k = 0$$
   e quindi
   $$0 \leq e_{k+1} \leq \lambda e_k \leq \lambda\lambda e_{k-1} \leq \cdots \leq \lambda^{k+1} e_0$$
   per il teorema dei due carabinieri tende a zero
   $$\lim_k e_k = 0$$
3. Per assurdo sia $\beta \neq \alpha$ tale che $\beta = g(\beta)$. Abbiamo $|\alpha - \beta| > 0$. Quindi $|g(\beta) - g(\alpha)| > 0$. Per il teorema di Lagrange ciò è uguale a
   $$\underbrace{|g'(\xi)|}_{<1} \cdot |\beta - \alpha| < |\beta - \alpha|$$
   che è assurdo

Possiamo rilassare l'ipotesi escludendo α dall'intervallo. Per avere una contrazione non è necessaria C^1, basta che la funzione di Lipschitz con $L < 1$ (ulteriore rilassamento).

Se la convergenza è monotona (che non sappiamo), potrebbe considerare solo un intervallo sinistro o destro. Sia infatti $-1 < g'(x) < 0$. Disegnino costante di Dottie convergenza. Salto a destra e a sinistra,

quindi in questo caso non poso prendere l'intervallo solo destro o solo sinistro (convergenza alternata). Se invece non cambia segno $g'(x) < -1$ possiamo considerare solo una parte dell'intervallo in quanto la convergenza è monotona. Se invece $g'(x) > 1$ abbiamo una divergenza.

Vediamo ora come scegliere x_0. Se $0 < g'(x) < 1$, quindi convergenza monotona, possiamo scegliere $x_0 = a$ o $x_0 = b$ che sono gli estremi. Il problema sussiste quando $-1 < g'(x) < 0$ dove abbiamo una convergenza alternata nell'intervallo simmetrico $I_\delta(\alpha) \subseteq [a, b]$. Dobbiamo sapere qual'è il più vicino. Quando $0 < g'(x) < 1$ la convergenza è monotona

Come criterio di arresto (con tolleranza ε) abbiamo:

1. $$|x_{k+1} - x_k| < \varepsilon$$

2. $$|f(x_k)| < \varepsilon$$

Idealmente vorrei $e_k < \varepsilon$ ma per calcolare l'errore e_k serve α.

$$\begin{aligned} |x_{k+1} - x_k| &= |x_{k+1} - \alpha + \alpha - x_k| \\ &= |g(x_k) - g(\alpha) + \alpha - x_k| \\ &= |g'(\xi_k)(x_k - \alpha) - (x_k - \alpha)| \\ &= |1 - g'(\xi_k)| \cdot |x_k - \alpha| \\ e_k \leq & \underbrace{\frac{1}{|1 - g'(\xi_k)|}}_{\text{coefficiente di amplificazione}} \cdot \varepsilon \end{aligned}$$

- se $g'(\xi_k) < 0$ (convergenza alternata) allora $e_k < 1 \implies e_k \leq \varepsilon$
- se $g'(\alpha) \approx 0$ (convergenza veloce) allora $e_k \approx 1 \implies e_k \lesssim \varepsilon$