

1 Tracce di calcolo numerico

(gli esercizi contrassegnati da * e ** sono più impegnativi)

1.1 Sistema floating-point e propagazione degli errori

- la mantissa di un numero reale sta in $[0, 1]$, ma non è la sua parte frazionaria
- i numeri irrazionali hanno parte frazionaria (e mantissa) infinita
- la parte frazionaria di un numero razionale può essere finita o infinita a seconda della base
- si dia un'interpretazione geometrica del fatto che il massimo errore di arrotondamento ad m cifre dopo la virgola è la metà del massimo errore di troncamento
- ** si dimostri, usando le serie, che l'errore di arrotondamento ad m cifre dopo la virgola in base β è $\leq \beta^{-m}/2$
- si studi l'insieme dei numeri floating-point $\mathbb{F}(\beta, t, L, U) = \{\mu \in \mathbb{Q} : \mu = \pm(0.\mu_1\mu_2 \dots \mu_t)\beta^p, \mu_j \in \{0, 1, \dots, \beta - 1\}, p \in [L, U] \subset \mathbb{Z}\}$ (traccia: determinare cardinalità, estremi, variazione di densità, intorno di approssimazione, precisione di macchina, ...); quali sono i reali rappresentabili tramite questi numeri floating-point?
- si disegnino $\mathbb{F}(10, 1, -1, 1)$ e $\mathbb{F}(10, 2, -2, 2)$
- i numeri floating-point “grandi” (in modulo) sono interi e hanno moltissime cifre nulle; in un sistema floating-point con t cifre di mantissa in base β , i numeri interi con più di t cifre vengono arrotondati (se rappresentabili)
- la precisione di macchina, $\varepsilon_M = \beta^{1-t}/2$, non è il più piccolo numero floating-point positivo; che cos'è invece?
- si dimostri che vale anche $\varepsilon_M = \min\{\mu \in \mathbb{F}_+ : 1 + \mu > 1\}$; quali numeri floating-point si comportano come elemento neutro nell'addizione con un dato numero floating-point μ ?
- si dimostri che se un reale y approssima il reale x con un errore relativo $< 10^{-m}$, allora i due numeri hanno almeno m cifre decimali significative coincidenti, a meno che ... (traccia: si ragioni sulle mantisse)
- * detto ε_f l'errore relativo su una funzione (differenziabile) con variabili approssimate, si dia una veste più rigorosa alla “formula degli errori”

$$\varepsilon_{f(x,y)} \lesssim |x f'_x(x,y)/f(x,y)| \varepsilon_x + |y f'_y(x,y)/f(x,y)| \varepsilon_y$$

partendo dal caso di funzioni di una variabile e si applichi la formula alla stima dell'errore nelle operazioni aritmetiche (traccia: si utilizzi il teorema del valor medio)

- in Matlab si ha che $((1 + 10^{-15}) - 1)/10^{-15} = 1.110223024625157$, ma invece $((1 + 2^{-50}) - 1)/2^{-50} = 1$, dove $2^{-50} \approx 10^{-15}$; perché?
- si facciano esempi in cui la proprietà associativa non è valida in aritmetica di macchina per overflow oppure per effetto dell'arrotondamento
- la formula risolutiva classica per le equazioni di secondo grado perde precisione in aritmetica di macchina se $b^2 \gg 4|ac|$ oppure $b^2 \approx 4ac$: perché? si quantifichi la perdita di precisione nei due casi e si discutano possibili rimedi (traccia nel secondo caso **: quanta precisione si perderebbe approssimando con la soluzione doppia $-b/(2a)$? ...)
- * si trovino esempi di funzioni "instabili", cioè tali che per certi punti (x, y) si ha $\varepsilon_{f(x,y)} \gg \max\{\varepsilon_x, \varepsilon_y\}$ qualunque sia la rappresentazione di f , e di funzioni "stabili" per cui l'instabilità nel calcolo dipende dalla rappresentazione (iniziare da una variabile)
- riscrivere, se necessario, le seguenti espressioni per ovviare a possibili instabilità in aritmetica di macchina:

$$E_1(x) = x^5/(2 + x^4 + 0.1|x|) - x + 100(x^4 + 5)^{1/4}$$

$$E_2(x) = \sqrt{x^2 + 1} - |x| + 100x$$

$$E_3(x) = 3\sqrt{x^4 + 2} - (10x^7 + 1)/(x^5 + 1) + 10x^2$$

- * la formula di ricorrenza in avanti $I_n = 1 - nI_{n-1}$, $n = 1, 2, \dots$, $I_0 = 1 - e^{-1}$, soddisfatta da $I_n = e^{-1} \int_0^1 x^n e^x dx$, è fortemente instabile in aritmetica di macchina (perché?), ma può essere stabilizzata usandola all'indietro; si dimostrino convergenza e stabilità della formula all'indietro analizzando l'errore relativo

1.2 Complessità degli algoritmi numerici

- si dimostri che la complessità della formula di Laplace per il calcolo di un determinante è almeno $n!$ flops
- si dimostri che la complessità asintotica del metodo di eliminazione gaussiana è $2n^3/3$ flops (traccia: si incastrino la complessità tra due integrali)
- perché il calcolo di $\exp(x)$ (per $x > 0$) usando la formula $\exp(x) = (\exp(x/m))^m$, $m > [x]$ e la formula di Taylor per $\exp(x/m)$, è più efficiente dell'utilizzo diretto della formula di Taylor?

1.3 Soluzione numerica di equazioni non lineari

- detto e_n l'errore assoluto del metodo di Newton per $f(x) = 0$, in ipotesi di convergenza e per $f \in C^2(I)$, dove I è un intervallo chiuso e limitato contenente

la radice in cui $f'(x) \neq 0$, si dimostri la disuguaglianza $e_{n+1} \leq Ke_n^2$ con K costante opportuna

approfondimento *: partendo da tale disuguaglianza, si arrivi ad enunciare e dimostrare un risultato di convergenza locale

- si giustifichi il fatto che nel calcolo di $\sqrt{2}$ risolvendo l'equazione $x^2 - 2 = 0$ nell'intervallo $(1, 2)$, il metodo di bisezione guadagna in media 1 cifra decimale significativa ogni 3-4 iterazioni, mentre il metodo di Newton raddoppia il numero di cifre significative ad ogni iterazione; questo comportamento vale in generale per entrambi i metodi? (traccia: si ragioni sull'errore relativo)
- si dimostri che in ipotesi di convergenza il numero di iterazioni che garantiscono un errore assoluto $\leq \varepsilon$ (dove ε è una tolleranza) è $n \geq c_1 \log(\varepsilon^{-1}) + c_2$ per il metodo di bisezione e $n \geq c_3 \log(\log(\varepsilon^{-1})) + c_4$ per il metodo di Newton, dove c_1, c_2, c_3, c_4 sono opportune costanti
- perché la quantità $|x_{n+1} - x_n|$ è una buona stima dell'errore assoluto $|x_n - \xi|$ per il metodo di Newton (almeno per n abbastanza grande)?
- si studino numericamente le seguenti equazioni “storiche”:
 - $x^3 - 2x - 5 = 0$ (su questa equazione Newton illustrò il suo metodo)
 - $m = x - E \sin x$ (equazione di Keplero: m è l'anomalia media di un pianeta, E l'eccentricità della sua orbita; si assuma ad esempio $m = 0.8, E = 0.2$)

(traccia: isolare gli zeri anche graficamente, verificare l'applicabilità dei metodi di bisezione e di Newton, ...)

- se si applica il metodo di Newton con $x_0 \neq 0$ all'equazione $\text{sign}(x)|x|^{1/2} = 0$ si ottiene una successione “stazionaria”, mentre per $\text{sign}(x)|x|^{1/3} = 0$ la successione è addirittura divergente; dove sta il problema?
- ** abbiamo visto che il metodo di Newton converge globalmente se $f \in C^2[a, b]$, $f(a)f(b) < 0$, $f''(x) > 0$ oppure $f''(x) < 0$ in $[a, b]$ e $f(x_0)f''(x_0) > 0$. Si dimostri che il metodo converge anche nelle seguenti ipotesi, per qualsiasi scelta di $x_0 \in [a, b]$: $f \in C^2[a, b]$, $f(a)f(b) < 0$, $f'(x) \neq 0$ e $f''(x) \geq 0$ oppure $f''(x) \leq 0$ in $[a, b]$, $|f(a)/f'(a)| < b - a$ e $|f(b)/f'(b)| < b - a$ (traccia: ci sono quattro situazioni geometriche possibili, dove può cadere la prima iterata x_1 ? ...)
- si studi la soluzione delle equazioni $5x - e^{-x} = 0$ e $x - e^{-x} = 0$ con le iterazioni di punto fisso (per la seconda si isoli la soluzione col teorema degli zeri per applicare un risultato di convergenza locale)
- perché il metodo di Newton si può interpretare come iterazione di punto fisso? quando ci si aspetta ordine di convergenza $p > 2$? (traccia: si ricordi che un'iterazione di punto fisso ha ordine di convergenza p se e solo se ...)

1.4 Interpolazione e approssimazione di dati e funzioni

- * si ricavi la forma di Lagrange dell'errore di interpolazione di grado n per $f \in C^{n+1}[a, b]$ (si noti l'analogia con il resto della formula di Taylor), $E_n(x) = f(x) - \Pi_n(x) = f^{(n+1)}(\eta) \omega(x)/(n+1)!$, dove $\omega(x) = (x - x_0) \dots (x - x_n)$, $\eta \in \text{int}(x, x_0, \dots, x_n)$ (traccia: si utilizzi la funzione ausiliaria $G(u) = E_n(u) - \omega(u)(E_n(x)/\omega(x))$ e si applichi il teorema di Rolle)
- * partendo dalla forma di Lagrange dell'errore di interpolazione polinomiale di grado n per $f \in C^{n+1}[a, b]$, si dimostri che il massimo errore nel caso di nodi equispaziati è $\leq \max_{x \in [a, b]} \{|f^{(n+1)}(x)|\} h^{n+1}/(4(n+1))$ dove $h = (b - a)/n$; perché da questa stima non si può dedurre la convergenza per $f \in C^\infty[a, b]$? (vedi controesempio di Runge, $f(x) = 1/(1 + x^2)$, $x \in [-5, 5]$)
- si dimostri che l'interpolazione quadratica a tratti converge uniformemente per $f \in C^3[a, b]$ con un errore $\mathcal{O}(h^3)$, dove $h = \max_i \Delta x_i$; c'è qualche vincolo sul numero di nodi? si utilizzi poi la stima trovata per decidere a priori quanti nodi usare nell'interpolazione quadratica a passo costante di $f(x) = \sin x$ in $[0, \pi]$ per avere un errore $< 10^{-8}$ (traccia: si utilizzi localmente la stima dell'errore di interpolazione polinomiale a passo costante)
- si dimostri che le seguenti affermazioni riguardo una formula di quadratura $\sum_{i=0}^n w_i f(x_i) \approx \int_a^b f(x) dx$, $\{x_i\}$ nodi distinti in $[a, b]$, sono equivalenti:
 - la formula è "interpolatoria" (cioè ottenuta integrando il polinomio interpolatore sui nodi)
 - $w_i = \int_a^b \ell_i(x) dx$, $i = 0, \dots, n$
 - la formula è esatta sui polinomi di grado $\leq n$
- * si dimostri che, per $f \in C^{s+1}[a, b]$, le formule di quadratura "composte" ottenute integrando un'interpolante polinomiale a tratti di grado locale fissato s costruita su un campionamento $\{(x_i, f(x_i))\}$, $i = 0, \dots, n$, sono convergenti con un errore che è almeno $\mathcal{O}(h^{s+1})$, dove $h = \max \Delta x_i$ (c'è qualche vincolo sul numero di nodi?)
- * si verifichi che, come le formule interpolatorie, le formule di quadratura composte hanno tutte la forma di somma pesata $\sum_{i=0}^n w_i f(x_i)$, dove i $\{w_i\}$ sono opportuni pesi (traccia: tali formule sono localmente "interpolatorie", cioè ottenute integrando un singolo polinomio interpolatore di grado s in $[x_{ks}, x_{(k+1)s}]$, $k = 0, \dots, (n : s) - 1$)
- si ricavi la formula di quadratura composta di grado 2, detta di Cavalieri-Simpson (o delle parabole), nel caso di campionamento a passo costante
- data una formula di approssimazione $\phi(h)$ di una quantità ϕ_0 , con la struttura $\phi(h) = \phi_0 + ch^p + \mathcal{O}(h^q)$, $q > p > 0$, utilizzando opportune combinazioni lineari di $\phi(h)$ e $\phi(h/2)$ si ricavino:

- una stima a posteriori della parte principale dell'errore
- una nuova approssimazione con errore di ordine $\mathcal{O}(h^q)$ (estrapolazione)

*approfondimento ***: come si potrebbe iterare il procedimento di estrapolazione per $\phi(h) = \phi_0 + c_1 h^{p_1} + c_2 h^{p_2} + \dots + c_m h^{p_m} + \mathcal{O}(h^q)$, $0 < p_1 < p_2 < \dots < p_m < q$? (traccia: si utilizzi una sequenza di passi $h, h/2, h/4, \dots$)

- si verifichi che il rapporto incrementale standard $\delta_+(h) = (f(x+h) - f(x))/h$ e quello "simmetrico" $\delta(h) = (f(x+h) - f(x-h))/(2h)$ hanno la struttura dell'esercizio precedente per f sufficientemente regolare (traccia: si utilizzi la formula di Taylor; in particolare, chi è q per $\delta(h)$ con $f \in C^5$)? si applichi l'esercizio precedente a vari esempi, controllando l'accuratezza delle approssimazioni e delle stime dell'errore
- perché nella derivazione numerica con $\delta_+(h)$ conviene prendere un passo dell'ordine di $\sqrt{\varepsilon}$, dove ε è il massimo errore su f ? e nel caso si usi $\delta(h)$?
- si rifletta sull'instabilità intrinseca dell'operazione funzionale di derivazione (fare un esempio) e sul fatto che l'operazione di integrazione invece è stabile; perché l'effetto dell'instabilità viene mitigato usando formule di derivazione numerica con errore di ordine maggiore? ad es. con $\delta(h)$ invece di $\delta_+(h)$, oppure applicando l'estrapolazione
- * gli esercizi precedenti mostrano che, in presenza di rumore, per calcolare la derivata è ragionevole non utilizzare tutti i dati di un campionamento fitto, ma campionare con un passo ottimale: individuare un criterio pratico per la determinazione di tale passo nel caso di $\delta_+(h)$, utilizzando un'approssimazione ai minimi quadrati
- detta $T(h)$ la formula di quadratura composta dei trapezi a passo costante e sapendo che $T(h) = \int_a^b f(x) dx + ch^2 + \mathcal{O}(h^4)$ per $f \in C^4[a, b]$, si calcolino stima a posteriori dell'errore e formula di estrapolazione su vari esempi di funzioni integrabili elementarmente, confrontando con l'errore effettivo
- ** abbiamo visto che la "risposta alle perturbazioni" su f dell'interpolazione polinomiale è legata alla quantità $\Lambda_n = \max_{x \in [a, b]} \sum_{i=0}^n |\ell_i(x)|$ (costante di Lebesgue), dove gli $\{\ell_i\}$ sono i polinomi elementari di Lagrange; quale quantità potrebbe svolgere un ruolo analogo nel caso delle formule di quadratura interpolatorie e composte, e perché in particolare le formule con pesi positivi sono sicuramente stabili? (traccia: si utilizzi il fatto che le formule interpolatorie e composte danno risultato esatto sulle funzioni costanti)