introduzione
ai
METODI NUMERICI

Giovanni Mancarella

Indice

Capitolo 1

INTRODUZIONE

RAPPRESENTAZIONE DI VALORI NUMERICI COL CALCOLATORE

1.1 Introduzione

1.2 Rappresentazione binaria di numeri interi

Nella notazione posizionale araba un numero intero è rappresentato da una sequenza di cifre ed ogni cifra rappresenta un valore diverso a seconda della posizione che essa occupa nella sequenza; così il valore numerico rappresentato dalla sequenza 123 in base 10 è dato da

$$3 \cdot 10^0 + 2 \cdot 10^1 + 1 \cdot 10^2 \tag{1.1}$$

Per rappresentare un numero in base n sono necessari n simboli distinti; maggiore è n, più compatta sarà la rappresentazione del numero, ma sarà necessario utilizzare un maggior numero di simboli.

Per ragioni di semplicità si è scelto per i calcolatori la base 2; sono quindi necessari due soli simboli, che possono essere rappresentati da due diversi stati di un circuito logico (detto bit); questi due simboli sono convenzionalmente indicati dallo 0 e dall'1, con gli stessi valori numerici che essi hanno nella rappresentazione decimale. Così il numero binario 1101011 rappresenta il valore numerico:

$$1 \cdot 2^0 + 1 \cdot 2^1 + 1 \cdot 2^3 + 1 \cdot 2^5 + 1 \cdot 2^6 \tag{1.2}$$

È evidente che se si hanno a disposizione n bit si possono rappresentare tutti gli interi compresi tra 0 e $2^n - 1$. Di fatto in un calcolatore viene utilizzato un numero fissato (e ovviamente finito) di bit per la rappresentazione di qualsiasi numero (la scelta più comune è $n = 32$) e quindi l'intervallo di valori che possono essere rappresentati è finito; in rari casi viene fatta la scelta di rappresentare solo valori positivi, più comunemente quella di rappresentare anche valori negativi; in tal caso si sceglie di rappresentare con n bit i numeri naturali (che nel linguaggio dei calcolatori sono chiamati sempre interi) compresi tra -2^{n-1} e $2^{n-1} - 1$. Con 32 bit a disposizione questo intervallo va da $\sim -10^9$ a $\sim 10^9$.

Per rappresentare i numeri naturali una possibile scelta sarebbe di utilizzare uno dei bit a disposizione per rappresentare il segno e gli altri per rappresentare il valore assoluto del numero.

Si preferisce invece di utilizzare la rappresentazione cosiddetta del *complemento a zero* . Facciamo un esempio: avendo a disposizione tre soli bit potremo rappresentare i numeri da -4 a 3 nel modo seguente:

$$3 \equiv 011$$

$$\begin{aligned} 2 &\equiv 010 \\ 1 &\equiv 001 \\ 0 &\equiv 000 \\ -1 &\equiv 111 \\ -2 &\equiv 110 \\ -3 &\equiv 101 \\ -4 &\equiv 100 \end{aligned}$$

Ogni numero si ottiene dal precedente sottraendogli uno e senza preoccuparsi del possibile riporto sul quarto *bit*. (*complemento a zero* vuol dire che applicando le regole della somma alle rappresentazioni di n e $-n$ si ottiene, limitandosi a tre *bit* e tralasciando il riporto, sempre zero).

In questo modo è possibile effettuare somme e sottrazioni nello stesso modo. Ad esempio:

$$\begin{aligned} 1-3 \equiv \quad & 001 \quad + \\ & 101 \quad = \\ \\ & 110 \quad \equiv -2 \end{aligned}$$

1.3 Rappresentazione di numeri reali

Per rappresentare al calcolatore numeri 'con la virgola' si utilizza sempre un numero finito di bit; il numero di cifre sarà dunque finito e quello che nel linguaggio dei calcolatori viene chiamato numero $reale$ è di fatto un numero $razionale$. Viene utilizzata la notazione scientifica, per cui un numero x viene scritto come:

$$x = s \cdot m \cdot 2^c \tag{1.3}$$

dove s vale -1 o 1 e rappresenta il segno di x, m è compreso tra 0 e 1 e c è scelto in modo tale che la prima cifra dopo la virgola di m non sia nulla.

Dei bit a disposizione (supponiamo che siano 32) uno viene utilizzato per il segno (bastano i due valori che può assumere un bit per distinguere i due casi -1 e 1) e gli altri vengono suddivisi tra la rappresentazione di m e quella di c. Supponiamo ad esempio di utilizzare 8 bit per rappresentare c: con questi possiamo rappresentare 255 valori distinti che utilizziamo per rappresentare i valori numerici interi compresi tra -128 e 127. L'intervallo di potenze che è possibile rappresentare va da:

$$2^{-128} \simeq 0.29 \cdot 10^{-38} \tag{1.4}$$

a:

$$2^{127} \simeq 1.7 \cdot 10^{38} \tag{1.5}$$

I restanti 23 bit li utilizzeremo per rappresentare m. Con 23 bit possiamo rappresentare tutti gli interi positivi compresi tra 0 e $2^{23}-1$, e li utilizzeremo per rappresentare nella notazione binaria l'intero più vicino a $m \cdot (2^{23}-1)$; in tal modo avremo la rappresentazione di m con una precisione data da:

$$p \simeq \frac{1}{2^{23}} \simeq 10^{-7} \tag{1.6}$$

ossia con circa 7 cifre (decimali) significative.

Oltre alla rappresentazione a 32 bit (detta in *precisione semplice*) sono utilizzate quelle a 64 (*precisione doppia*) e a 128 (*precisione quadrupla*). Inoltre la scelta del numero di bit da utilizzare per la rappresentazione di m e di c può essere diversa rispetto a quella descritta in questo paragrafo e varia a seconda del calcolatore e del linguaggio di programmazione utilizzato.

1.4 Errori di troncamento

1.5 Calcolo approssimato della derivata di una funzione.

Dalla formula di Taylor:

$$f(x+h) = f(x) + f'(x)\cdot h + \frac{1}{2}f''(\xi)\cdot h^2 \qquad , \qquad \xi \in [x, x+h] \tag{1.7}$$

si ottiene:

$$f'(x) = \frac{f(x+h) - f(x)}{h} + o(h) \tag{1.8}$$

ma se riscriviamo la formula di Taylor per $f(x+h)$ e $f(x-h)$:

$$f(x+h) = f(x) + f'(x)\cdot h + \frac{1}{2}f''(x)\cdot h^2 + \frac{1}{6}f'''(\xi)\cdot h^3 \qquad , \qquad \xi \in [x, x+h] \tag{1.9}$$

$$f(x-h) = f(x) - f'(x)\cdot h + \frac{1}{2}f''(x)\cdot h^2 - \frac{1}{6}f'''(\eta)\cdot h^3 \qquad , \qquad \eta \in [x-h, x] \tag{1.10}$$

e sottraiamo membro a membro queste due espressioni, si avrà:

$$f'(x) = \frac{f(x+h) - f(x-h)}{2\cdot h} + o(h^2) \tag{1.11}$$

Come si vede, il rapporto incrementale simmetrico in (1.11) fornisce,a parità di calcoli, una approssimazione alla derivata della funzione migliore rispetto a quello asimmetrico in (1.8).

```
c                  DERIVATA.FOR
c calcolo della derivata della funzione sqrt(x) al variare del passo

        program derivata
implicit real*8 (a-h,o-z)
        character*30,char1,char2
        data char1/'Immetti il punto:          '/
        data char2/'Immetti il passo iniziale: '/

        write(5,100) char1
        read(6,200) x
        write(5,100) char2
        read(6,200) h

        ex = .5 / sqrt(x)
        print * ,'valore ''esatto''', ex
        print *,'   passo                    valore appr.
     >errore rel. '
        der = 100.
        do while (der.ne.0.)
          x2 = x+h*.5
          x1 = x-h*.5
          der = ( f(x2) - f(x1) ) / h
          print * , h , der , ( der - ex ) / ex
          h   = h * .3
        enddo
        stop
100     format('$',a)
200     format(e10.3)
        end

real*8 function f(x)
implicit real*8 (a-h,o-z)
f=sqrt(x)
return
```

```
end


>  run derivata
Immetti il punto:            5.
Immetti il passo iniziale:    2.
valore 'esatto'  0.223606797749979
   passo                    valore appr.             errore rel.
  2.00000000000000       0.224744871391589       5.089620052081282E-003
 0.600000000000000       0.223707579627312       4.507102572355277E-004
 0.180000000000000       0.223615855109231       4.050574196757887E-005
 5.400000000000000E-002  0.223607612807157       3.645046511361108E-006
 1.620000000000000E-002  0.223606871104280       3.280504053178632E-007
 4.860000000000000E-003  0.223606804351823       2.952434523410666E-008
 1.458000000000000E-003  0.223606798344314       2.657945175574127E-009
 4.374000000000000E-004  0.223606797802756       2.360271828733998E-010
 1.312200000000000E-004  0.223606797756391       2.867637263769756E-011
 3.936600000000000E-005  0.223606797751879       8.496224886073468E-012
 1.180980000000000E-005  0.223606797751879       8.496224886073468E-012
 3.542939999999999E-006  0.223606797714275      -1.596715075418505E-010
 1.062882000000000E-006  0.223606797588931      -7.202304501326532E-010
 3.188645999999999E-007  0.223606797031842      -3.211604286866102E-009
 9.565937999999996E-008  0.223606800745763       1.339755487294364E-008
 2.869781399999999E-008  0.223606803840697       2.723852075670061E-008
 8.609344199999995E-009  0.223606819315368       9.644335029961220E-008
 2.582803259999999E-009  0.223606750539055      -2.111336693394317E-007
 7.748409779999996E-010  0.223606693225462      -4.674478524547194E-007
 2.324522933999999E-010  0.223607075316086       1.241313367941485E-006
 6.973568801999996E-011  0.223605801680671      -4.454557366257399E-006
 2.092070640599999E-011  0.223607924406363       5.038560524032700E-006
 6.276211921799996E-012  0.223593772901749      -5.824889207773579E-005
 1.882863576539999E-012  0.223593772901749      -5.824889207761167E-005
 5.648590729619996E-013  0.223279295021437      -1.464636727672453E-003
 1.694577218885999E-013  0.222755165220917      -3.808616453663814E-003
 5.083731656657996E-014  0.227122913558582       1.572454792959765E-002
 1.525119496997399E-014  0.232946578008802       4.176876710727965E-002
 4.575358490992197E-015  0.291183222511003       0.302210958884100
 1.372607547297659E-015  0.323536913901114       0.446901065426777
 4.117822641892977E-016  0.000000000000000E+000  -1.00000000000000
```

Capitolo 2

ZERI DI FUNZIONI

2.1 Introduzione

Il problema è quello della ricerca di una soluzione della equazione:

$$f(x) = 0 \tag{2.1}$$

discuteremo due classi di metodi: in un caso (metodo di $bisezione$) si suppone che f sia continua in un intervallo contenente lo zero. Nell'altro (metodi del $punto\ fisso$,di $Newton$,della $secante$) supporremo che essa sia anche derivabile.

2.2 Metodo di bisezione

Si parte da un intervallo $[a, b]$ per cui si abbia $f(a) \cdot f(b) < 0$; questo intervallo va determinato o a partire da proprietà generali note della funzione oppure calcolando direttamente i valori della funzione in vari punti fino a trovare una coppia di punti che abbia la proprietà richiesta. La continuità della funzione ci assicura che in $[a, b]$ ci sarà certamente almeno uno zero. Per cercarlo, calcoliamo il valore della funzione in $\frac{a+b}{2}$; a seconda del suo segno possiamo stabilire se lo zero si deve trovare in $[\frac{a+b}{2}, b]$ o in $[a, \frac{a+b}{2}]$. A questo punto si ripete la procedura descritta applicandola al nuovo intervallo di ampiezza $\frac{b-a}{2}$, e così via finchè l'ampiezza dell'intervallo non è inferiore alla precisione con cui si vuole conoscere il valore dello zero.

E' evidente che la strategia di bisezione dell'intervallo e' la migliore che si puo' adottare quando non si conosca nessuna proprieta' della funzione oltre alla continuita': lo zero ha la stessa probabilita' di trovarsi a destra o a sinistra del punto di mezzo.

Dopo N calcoli della funzione, abbiamo la garanzia di conoscere la posizione di uno zero della funzione con un precisione $\frac{b-a}{2^N}$. Si tratta di una convergenza non molto rapida, ma sicura. I metodi che discuteremo in seguito utilizzano la proprieta' di derivabilita' della funzione. Sono piu' rapidi nella convergenza verso lo zero, ma necessitano di un punto di partenza che gli sia sufficientemente vicino; altrimenti possono anche divergere. In questo senso chiameremo $locali$ questi ultimi metodi e $non\text{-}locale$ il metodo di $bisezione$. Una possibile strategia ottimale puo' essere quella di utilizzare il metodo di $bisezione$ per i primi passi e successivamente un metodo $locale$.

2.3 Metodo del punto fisso

La (2.1) puo' essere sempre, ed in infiniti modi, posta nella forma:

$$x = g(x) \tag{2.2}$$

Partendo da un x_0 arbitrario, la successione definita dalla relazione di ricorrenza:

$$x_{n+1} = g(x_n) \tag{2.3}$$

puo', sotto certe condizioni, convergere ad una soluzione della (2.2); un'insieme di condizioni che assicurano la convergenza sono le seguenti ($Teorema\ del\ punto\ fisso$):

1. dato un intervallo $I = [a, b]$ in cui $g(x)$ e' definita, $g(x) \in I \ \ \forall \ \ x \in I$.
2. $g(x)$ e' derivabile in I ed $\exists k, 0 < k < 1$, tale che $|g'(x)| \leq k \ \ \forall \ \ x \in I$.
3. $x_0 \in I$.

(potete riconoscere, in una forma meno generale, il Teorema delle contrazioni).

Queste condizioni assicurano l'esistenza e l'unicita' di uno $\xi \in I$ che e' soluzione della (2.2) e che la successione $x_0, x_1, x_2, ...$ definita dalla (2.3) tende a ξ.

L'esistenza si dimostra osservando che, per la prima ipotesi, si ha: $g(a) \geq a, g(b) \leq b$; quindi la funzione $h(x) = x - g(x)$ e' negativa in a e positiva in b. Per la continuita' di $g(x)$, $h(x)$ deve dunque avere uno zero in I che sara' soluzione della (2.2).

Sia ora $e_n = \xi - x_n$; si ha:

$$e_n = \xi - x_n = g(\xi) - g(x_{n-1}) = g'(\eta_n) \cdot (\xi - x_{n-1}) = g'(\eta_n) \cdot e_{n-1} \Rightarrow |e_n| \leq k \cdot |e_{n-1}| \ \ \forall n \tag{2.4}$$

dove η_n e' un punto compreso tra ξ ed x_{n-1}. Possiamo dunque scrivere:

$$|e_n| \leq k \cdot |e_{n-1}| \leq k^2 \cdot |e_{n-2}| \leq k^3 \cdot |e_{n-3}| \ \ ... \ \ \leq k^n \cdot |e_0| \ \ \forall n \tag{2.5}$$

allora:

$$k < 1 \Rightarrow \ \lim_{n\to\infty} |e_n| = 0 \ \Rightarrow \lim_{n\to\infty} x_n = \xi \tag{2.6}$$

per dimostrare l'unicita' osserviamo che se ci fosse un altro zero $\eta \neq \xi$ si avrebbe, ponendo $x_0 = \eta$:

$$x_1 = g(x_0) = g(\eta) = \eta \Rightarrow e_0 = \xi - \eta \ \ , \ \ e_1 = \xi - x_0 = \xi - \eta = e_0 \tag{2.7}$$

ma l'uguaglianza di e_1 ed e_0 e' incompatibile con l'ipotesi $k < 1$ (vedi (2.5)), a meno che non si abbia $e_0 = 0$ e quindi $\eta = \xi$.

Le condizioni del teorema del punto fisso possono essere difficilmente verificabili. Una formulazione piu' utile per le applicazioni e' la seguente:

- Se ξ e' una soluzione della (2.2) , se $g(x)$ e' derivabile con derivata continua in un intervallo aperto contenente ξ e se $|g'(\xi)| < 1$ allora esiste un ε tale che la successione definita da (2.3) tende a ξ per qualunque punto iniziale $x_0 \in [\xi - \varepsilon, \xi + \varepsilon]$.

si vede facilmente che nelle ipotesi enunciate ci si puo' ricondurre a quelle del teorema del punto fisso.

Scritto in questa forma, il teorema evidenzia bene quale e' la difficolta' nell'applicazione di questi metodi che abbiamo chiamato $locali$: non possiamo stabilire a priori quanto il punto di partenza deve essere vicino a ξ perche' si abbia la convergenza (a meno, ovviamente, di non calcolare la derivata della funzione in un certo intervallo).

Notiamo che la condizione $|g'(\xi)| < 1$ puo' essere realizzata in vari modi, sfruttando le diverse possibilita' di passare dalla (2.1) alla (2.2). Nelle applicazioni non e' evidentemente possibile verificare se essa e' relizzata (non conoscendo il valore di ξ). Quello che si fa e' verificare direttamente, calcolandone un certo numero di termini, se la successione degli x_i e' 'verosimilmente' convergente o no.

Il metodo di $Newton$, comunque, fornisce una trasformazione sulla funzione che da' $sempre$ $|g'(\xi)| < 1$, e per di piu' converge piu' rapidamente di un generico metodo di punto fisso.

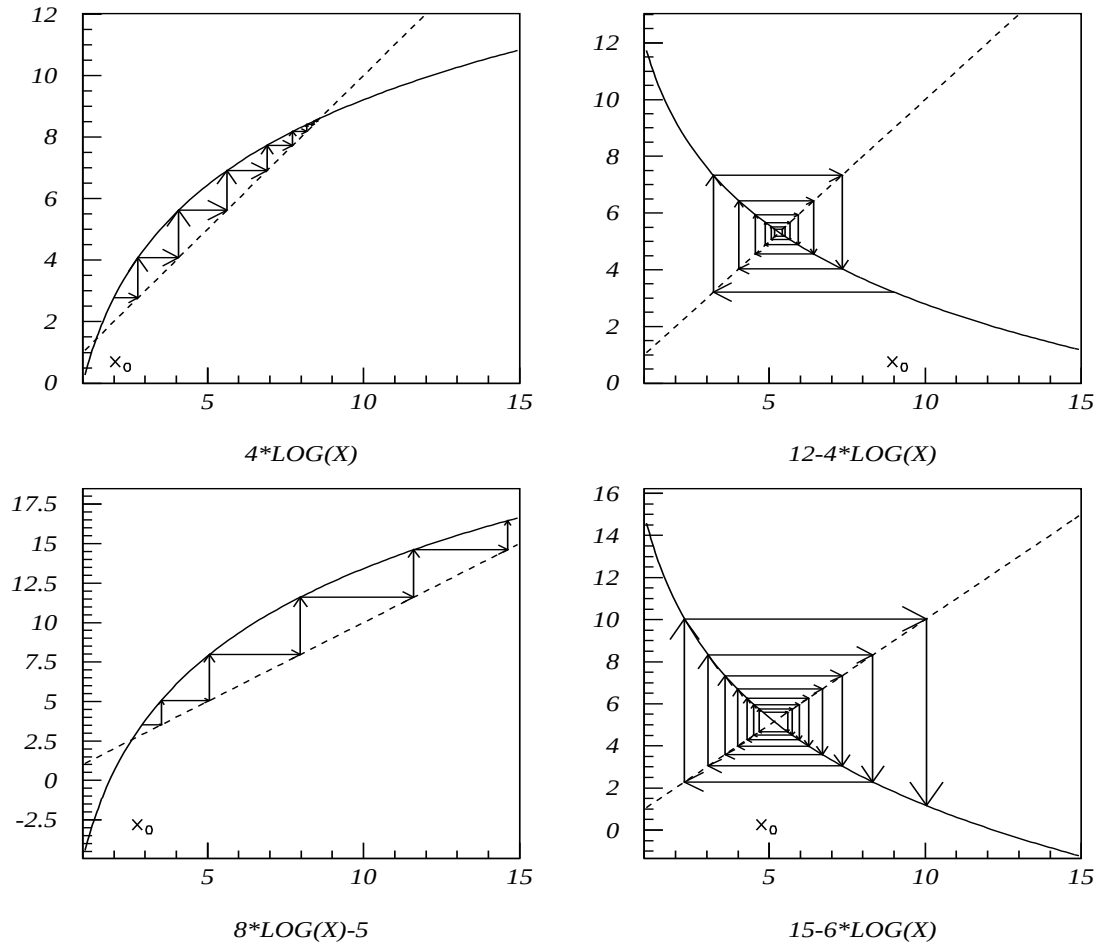


Figura 2.1: Illustrazione grafica del metodo del punto fisso

2.3.1 Esempio

L'equazione:

$$\sin(5 \cdot x) = 0 \tag{2.8}$$

può essere trasformata in una equazione del tipo (2.2) in vari modi:

$$g(x) = x + \sin(5 \cdot x) \tag{2.9}$$

$$g(x) = x + \sin(5 \cdot x) \tag{2.10}$$

$$g(x) = x + \frac{1}{10} \cdot \sin(5 \cdot x) \tag{2.11}$$

$$g(x) = x + \frac{1}{5} \cdot \sin(5 \cdot x) \tag{2.12}$$

Nel primo caso la derivata di $g(x)$ negli zeri della (2.8) è sempre maggiore di 1 in valore assoluto. Nel secondo in metà di questi zeri è minore di 1; nel terzo, sempre in metà degli zeri,è nulla.

Quindi il metodo del punto fisso non convergerà nel primo caso, convergerà nel secondo e convergerà più rapidamente nel terzo. Nella tabella seguente sono riportati i valori numerici ottenuti nei tre casi; la soluzione ottenuta corrisponde alla soluzione $\xi = 3 \cdot \pi \simeq 9.42477796076938$.

```
punto iniziale :    10.0000000000000
   1    9.73762514629607        9.97376251462961        9.94752502925921
   2    8.73764666563747        9.93512761972641        9.84678833825099
   3    8.44780267983866        9.87950445096901        9.67517010090217
   4    7.46264178981871        9.80320324773716        9.48524989137973
   5    7.08624600553690        9.70832163063660        9.42569516247791
   6    6.31952948954235        9.60949098786847        9.42477796398439
   7    6.50025190526501        9.52971534883983        9.42477796076938
   8    7.38471074813210        9.47962114062895        9.42477796076938
   9    6.68459606167351        9.45254191910047        9.42477796076938
  10    7.59093548646842        9.43870448344550        9.42477796076938
  11    7.84372815469986        9.43174684787191        9.42477796076938
  12    8.84241426955097        9.42826310937514        9.42477796076938
  13    9.07017193096776        9.42652062326160        9.42477796076938
  14    10.0497922719676        9.42564930304095        9.42477796076938
  15    10.0332719259215        9.42521363328341        9.42477796076938
  16    9.93431133753400        9.42499579719867        9.42477796076938
  17    9.37469252842458        9.42488687900556        9.42477796076938
  18    9.62251034739545        9.42483241989016        9.42477796076938
  19    8.78721930071994        9.42480519033011        9.42477796076938
  20    8.74103508818667        9.42479157554979        9.42477796076938
  21    8.46744677745182        9.42478476815959        9.42477796076938
  22    7.47020329902473        9.42478136446448        9.42477796076938
  23    7.12909529061203        9.42477966261693        9.42477796076938
  24    6.24341496105704        9.42477881169316        9.42477796076938
  25    6.04587114143845        9.42477838623127        9.42477796076938
  26    5.11878211549681        9.42477817350033        9.42477796076938
  27    5.56377771658346        9.42477806713485        9.42477796076938
  28    6.00363998189392        9.42477801395212        9.42477796076938
  29    5.01857919689640        9.42477798736075        9.42477796076938
  30    4.97874449516356        9.42477797406506        9.42477796076938
```

2.4 Metodo di Newton (o della tangente).

Il metodo di $Newton$ consiste nel partire da un punto x_0, nel calcolare lo zero della tangente in x_0 alla curva che rappresenta la funzione; sia x_1 questo zero. x_2 sara' poi lo zero della tangente in x_1, e cosi' via.

Si puo' scrivere una relazione di ricorrenza tra questi punti:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{2.13}$$

che e' del tipo (2.3) con:

$$g(x) = x - \frac{f(x)}{f'(x)} \tag{2.14}$$

quindi il metodo di $Newton$ e' un metodo di punto fisso che verifica le condizioni del secondo criterio del paragrafo (2.3); infatti:

$$g'(x) = \frac{f(x) \cdot f''(x)}{[f'(x)]^2} \tag{2.15}$$

se ξ e' uno zero semplice si avra' $g'(\xi) = 0$. Se invece ξ e' uno zero di ordine n:

$$f(\xi) \sim \alpha \cdot (x-\xi)^n \quad , \quad g'(\xi) \sim \frac{\alpha(x-\xi)^n \cdot \alpha n(n-1)(x-\xi)^{n-2}}{\alpha^2 n^2 (x-\xi)^{2(n-1)}} = \frac{n-1}{n} \tag{2.16}$$

$g'(\xi)$ e' ancora minore di 1 ma non nullo. Il fatto che $g'(\xi) = 0$ rende la convergenza piu' rapida che per il generico metodo di punto fisso; infatti, sviluppando $g(x_{n-1})$ intorno a ξ , si ha:

$$\begin{aligned} e_n &= \xi - x_n = g(\xi) - g(x_{n-1}) \\ &= g(\xi) - [g(\xi) + g'(\xi) \cdot (x_{n-1} - \xi) + \frac{1}{2} g''(\eta_n) \cdot (x_{n-1} - \xi)^2] \\ &= -\frac{1}{2} g''(\eta_n) \cdot {e^2}_{n-1} \end{aligned} \tag{2.17}$$

e quindi l'errore decresce quadraticamente invece che linearmente.

Se invece ξ e' uno zero di ordine n il metodo di $Newton$ resta convergente ma non quadraticamente. Se tuttavia si conosce n, utilizzando al posto della (2.14) la funzione:

$$g(x) = x - n \cdot \frac{f(x)}{f'(x)} \tag{2.18}$$

la convergenza quadratica e' ripristinata (esercizio!).

Un'osservazione finale: sembrerebbe da (2.17) che, se $|g''(\eta_n)| > 2$, il metodo non converge perche' $|e_n| > |e_{n-1}|$; il punto e' che, come detto piu' volte, bisogna partire da un punto sufficientemente vicino a ξ; se ad esempio:

$$-\frac{1}{2} \cdot g''(x) \simeq cost = k \quad per \quad x \simeq \xi \tag{2.19}$$

gli errori nelle successive iterazioni saranno:

$$e_0 \quad , \quad k \cdot {e_0}^2 \quad , \quad k^3 \cdot {e_0}^4 \quad , \quad k^7 \cdot {e_0}^8 \quad , \quad ... \tag{2.20}$$

e tenderanno a zero se $k \cdot e_0 < 1$.

2.5 Metodo della secante

Questo metodo sostituisce alla tangente del metodo di $Newton$ una secante alla curva. Sono necessari dunque due punti di partenza; ad ogni passo lo zero della secante attraverso due punti sostituisce il penultimo di questi. La relazione di ricorrenza e' la seguente:

$$x_{n+1} = \frac{x_{n-1} \cdot f(x_n) - x_n \cdot f(x_{n-1})}{f(x_n) - f(x_{n-1})} \tag{2.21}$$

Questo metodo presenta rispetto al metodo di $Newton$ due vantaggi:

- ad ogni passo e' necessario un solo calcolo della funzione

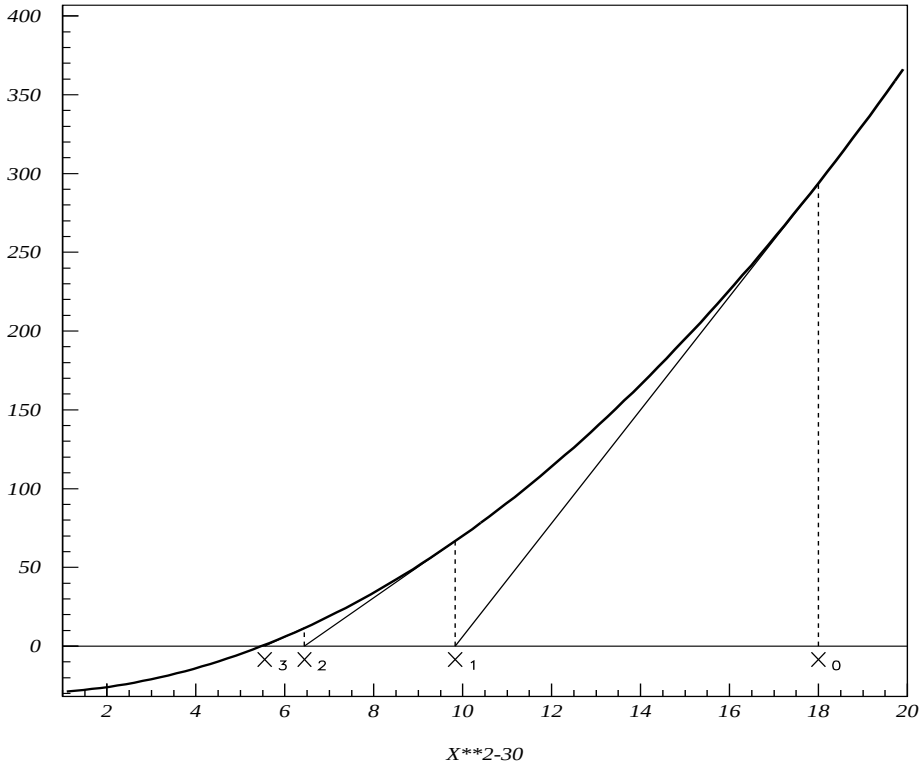


Figura 2.2: Illustrazione grafica del metodo di Newton

- non e' necessario alcun calcolo della derivata; di fatto, spesso nel metodo di $Newton$ il calcolo della derivata si effettua numericamente, il che equivale comunque a sostituire la tangente con una secante attraverso due punti vicini, ma con un numero maggiore di calcoli.

ed uno svantaggio, la convergenza e' piu' lenta:

$$e_n \sim e_{n-1} \cdot e_{n-2} \tag{2.22}$$

2.5.1 Esempio

Qui di seguito riportiamo il programma per l'applicazione del metodo di Newton e di quello della tangente alla equazione (2.8) ed i risultati che si ottengono; la prima colonna è relativa al metodo di Newton.

```
real*8 x,x1,x2,x3,xx,f2,f3
        print * , 'immettere il punto iniziale :  '
read *,x
        print * , 'punto iniziale : ',x
x1 = x                ! punto di partenza per il metodo di Newton
x2 = x                ! punti di partenza per il metodo della secante
x3 = x * 1.01d0       !
        f2 = sin(5.d0*x2)

        do while ( abs((x2-x3)/x2).gt.1.d-15 )
          x1 = x1 - tan(5.d0*x1) / 5.d0
          xx = x3
          f3 = sin(5.d0*x3)
          x3 = ( x2*f3 - x3*f2 ) / ( f3 - f2 )
          x2 = xx
          f2 = f3
          print * ,x1,x3
        enddo

        stop
        end
```

```
> run newton
immettere il punto iniziale :
9.6
punto iniziale :    9.60000000000000
  9.35997455137674        9.24676026115037
  9.42714526472731        9.44578109757738
  9.42477785020756        9.42212679786767
  9.42477796076938        9.42478222321864
  9.42477796076938        9.42477796064475
  9.42477796076938        9.42477796076938
  9.42477796076938        9.42477796076938
```

2.6 Zeri in piu' dimensioni

In piu' dimensioni, cioe' quando si ha un sistema di N equazioni in N incognite il problema fondamentale e' quello della localizzazione dello zero, cioe' di individuare una regione dello spazio che contenga lo zero. A partire da un punto di questa regione e' poi possibile innescare il metodo di $Newton$, che e' facilmente generalizzabile a piu' dimensioni.

Per la localizzazione dello zero bisogna far ricorso alle proprieta' delle funzioni che entrano nel sistema di equazioni.

Capitolo 3

METODI DI INTEGRAZIONE APPROSSIMATA

3.1 Introduzione

Il problema che intendiamo discutere e' quello di trovare un'espressione approssimata per l'integrale definito

$$I = \int_a^b f(x)\,dx \tag{3.1}$$

dove supponiamo, per il momento, che a e b siano finiti. Le espressioni che troveremo saranno tutte del tipo

$$I \simeq I_N = \sum_{i=1}^{N} \alpha_i f(x_i) \tag{3.2}$$

dove α_i sono costanti fissate, diverse per i vari metodi utilizzati, e x_i sono punti appartenenti all'intervallo $[a, b]$. Essi possono essere equispaziati o no.

In seguito discuteremo il metodo di $Simpson$, che utilizza punti equispaziati, e quello di $Gauss$, che richiede una particolare scelta di punti non equispaziati. Nel caso di una griglia equispaziata di punti, l'errore che si commette in (3.2) e' proporzionale ad una certa potenza, differente per in diversi metodi, dell'ampiezza dell'intervallo di base, e quindi inversamente proporzionale alla stessa potenza di N, il numero di punti della griglia. Anche per il metodo di $Gauss$ l'errore decresce al crescere di N ed in generale a parita' di numero di punti (cioe' di calcoli da effettuare) questo metodo fornisce un errore minore rispetto a quelli che utilizzano punti equispaziati.

Tuttavia questi ultimi risultano piu' vantaggiosi quando si imposta una procedura per il calcolo approssimato di I con un errore prefissato. Sia ε questo errore; una possibile strategia e' la seguente: Si calcola I_{N_1}, I_{N_2},I_{N_3}, ... con una suddivisione sempre piu' fitta, cioe' con $N_{i+1} > N_i$; si interrompe il calcolo quando

$$\left|\frac{I_{N_{i+1}} - I_{N_i}}{I_{N_i}}\right| < \varepsilon \tag{3.3}$$

ossia quando due valori successivi differiscono di meno della precisione richiesta. Si noti che abbiamo utilizzato ε come errore percentuale; questo e' sempre preferibile quando non si conosce a priori l'ordine di grandezza di I.

Ora, se si utilizza un metodo a griglia equispaziata e' possibile, scegliendo $N_{i+1} = 2 \cdot N_i$, riutilizzare, per il calcolo di $I_{N_{i+1}}$, i valori della funzione utilizzati per calcolare I_{N_i}, con un notevole risparmio di tempo di calcolo (il metodo di $Simpson$ richiede una suddivisione in un numero $dispari$

di punti; in tal caso, fra un passo e il successivo, va raddoppiato il numero di intervalli piuttosto che il numero di punti). Questo non e' possibile nel caso del metodo di $Gauss$ perche', come vedremo, due suddivisioni dell'intervallo che utilizzano un numero di punti diverso non hanno nessun punto in comune.

Discuteremo per ultimo il metodo $Monte\ Carlo$, che utilizza, invece delle proprieta' analitiche della funzione da integrare, le proprieta' statistiche delle medie di variabili casuali; e' un metodo che, per integrali multidimensionali, puo' essere competitivo coi metodi analitici.

3.2 Metodo di SIMPSON

Consideriamo una suddivisione dell'intervallo $[a, b]$ in un numero $pari$ di sotto-intervalli di ampiezza h; N sara' dunque $dispari$. Consideriamo un punto della griglia, x_i, ed i due punti adiacenti $x_i - h$ ed $x_i + h$.

Il metodo di $Simpson$ consiste nell'approssimare la funzione $f(x)$, in ciascuna delle coppie di intervalli elementari, con la parabola che coincide con essa nei tre punti $x_i - h$,x_i,$x_i + h$.

Per ottenere l'espressione di I_N e dell'errore, consideriamo lo sviluppo di Taylor della funzione $f(x)$ in un punto $x \in [x_i - h, x_i + h]$:

$$\begin{aligned} f(x) = f(x_i) \quad &+ \quad f^{'}(x_i) \cdot (x - x_i) + \frac{1}{2} \cdot f^{''}(x_i) \cdot (x - x_i)^2 + \frac{1}{3!} \cdot f^{'''}(x_i) \cdot (x - x_i)^3 + \\ &+ \quad \frac{1}{4!} \cdot f^{IV}(\xi) \cdot (x - x_i)^4 \quad , \quad \xi \in [x_i - h, x_i + h] \end{aligned} \tag{3.4}$$

se ora imponiamo che il polinomio

$$\overline{f}(x) = \alpha + \beta \cdot (x - x_i) + \frac{1}{2} \cdot \gamma \cdot (x - x_i)^2 \tag{3.5}$$

coincida con $f(x)$ nei tre punti anzidetti, otterremo che $\alpha = f(x_i)$ mentre β e γ ci forniranno due espressioni approssimate rispettivamente per $f^{'}(x_i)$ ed $f^{''}(x_i)$; in particolare:

$$f^{''}(x_i) \simeq \gamma = \frac{f(x_i - h) - 2 \cdot f(x_i) + f(x_i + h)}{h^2} \tag{3.6}$$

con un errore dato, in valore assoluto, da:

$$\frac{h^2}{12} \cdot |f^{IV}(\xi_1)| \quad , \quad \xi_1 \in [x_i - h, x_i + h] \tag{3.7}$$

(esercizio!).

Sostituendo ora $f(x)$ con $\overline{f}(x)$ nell'integrale otterremo l'espressione approssimata:

$$\int_{x_i - h}^{x_i + h} f(x)\, dx \simeq \int_{x_i - h}^{x_i + h} \overline{f}(x)\, dx = \frac{h}{3} \cdot (f(x_i - h) + 4 \cdot f(x_i) + f(x_i + h)) \tag{3.8}$$

l'errore commesso puo' essere stimato utilizzando la (3.4). L'integrale dei termini contenenti $f^{'}(x_i)$ ed $f^{'''}(x_i)$ e' nullo; quindi le sorgenti d'errore sono:

- la sostituzione di $f^{''}(x_i)$ con la (3.6). L'errore sull'integrale si calcola utilizzando la (3.7); si ottiene, sempre in valore assoluto:

$$|e_1| = \frac{h^5}{36} \cdot |f^{IV}(\xi_1)| \tag{3.9}$$

- il troncamento in (3.4) del termine del quarto ordine; l'errore e' dato da:

$$|e_2| = \frac{h^5}{60} \cdot |f^{IV}(\xi)| \tag{3.10}$$

quindi l'errore complessivo e' dato da:

$$e = e_1 + e_2 \leq |e_1| + |e_2| = \frac{h^5}{36} \cdot |f^{IV}(\xi_1)| + \frac{h^5}{60} \cdot |f^{IV}(\xi)| \leq \frac{4 \cdot h^5}{90} \cdot |f^{IV}(\overline{\xi})| \tag{3.11}$$

dove $\overline{\xi}$ sara' uno dei due punti ξ , ξ_1 .

Una valutazione piu' accurata dell'errore da' la stessa espressione ma con un fattore numerico $\frac{1}{90}$. Quello che e' importante in (3.11) e' l'andamento con h: il metodo di $Simpson$ e' di ordine h^5 sulla coppia di intervalli elementari.

Per l'integrale sull'intervallo $[a, b]$ dovremo sommare $\frac{N-1}{2}$ termini del tipo (3.8); l'espressione finale sara':

$$I_N = \frac{h}{3} \cdot [f(a) + 4 \cdot f(x_2) + 2 \cdot f(x_3) + 4 \cdot f(x_4) + ... + 2 \cdot f(x_{N-2}) + 4 \cdot f(x_{N-1}) + f(b)] \tag{3.12}$$

mentre l'errore globale puo' essere scritto:

$$e \leq \sum_{i=1}^{\frac{N-1}{2}} \left[\frac{4}{90} \cdot h^5 \cdot |f^{IV}(\overline{\xi}_i)| \right] = \frac{4}{90} \cdot h^5 \cdot \frac{N-1}{2} \cdot |f^{IV}(\eta)| = \frac{2}{90} \cdot h^4 \cdot (b-a) \cdot |f^{IV}(\eta)| \tag{3.13}$$

dove η e' ancora un punto interno all'intervallo $[a, b]$.

Come si vede, il metodo diventa di ordine h^4 sull'intervallo totale. Notiamo ancora che l'espressione ottenuta e' esatta per un polinomio del terzo ordine, che ha derivata quarta nulla, nonostante che la funzione sia stata apprrossimata da un polinomio del secondo ordine.Questo e' dovuto alla scelta del punto di mezzo x_i dell'intevallo $[x_i - h, x_i + h]$ come punto attorno al quale sviluppare la funzione (l'integrale dei termini contenenti le derivate prima e terza e' nullo).

3.2.1 Altri metodi con griglia equispaziata

Altri metodi piu' elementari approssimano la funzione con una funzione a gradini o con una poligonale ($metodo\ del\ trapezio$). Le espressioni che si ottengono sono sempre del tipo (3.2), ma gli errori sono di ordine piu' basso rispetto al metodo di $Simpson$.

3.3 Metodo di GAUSS

Il metodo di $Gauss$ utilizza le proprieta' dei polinomi ortogonali, che sono esposte in appendice. Esso permette di ottenere una espressione approssimata per l'integrale:

$$I = \int_a^b f(x) \cdot w(x)\, dx \tag{3.14}$$

dove $w(x)$ e' la funzione peso di un insieme di polinomi ortogonali sull'intervallo $[a, b]$. Indicheremo con $P_i(x)$ questi polinomi; i rappresenta il grado del polinomio.

Dimostreremo ora che il metodo di $Gauss$ permette di ottenere un'espressione per l'integrale che, utilizzando $n+1$ punti, e' esatta per qualunque polinomio di grado $2n+1$. Sia $p(x)$ un generico polinomio di grado $2n+1$; la divisione tra $p(x)$ e $P_{n+1}(x)$ da':

$$p(x) = q(x) \cdot P_{n+1}(x) + r(x) \tag{3.15}$$

dove $q(x)$ e' un polinomio di grado n ed $r(x)$ e' un polinomio di grado al piu' n. Utilizzeremo come punti della griglia gli $n+1$ zeri di $P_{n+1}(x)$; essi sono tutti semplici ed interni ad $[a,b]$. La (3.15) implica che in questi punti si avra':

$$r(x_i) = p(x_i) \quad , \quad i = 1, 2, ..., n+1 \tag{3.16}$$

sostituendo (3.15) in (3.14):

$$\int_a^b p(x) \cdot w(x)\, dx = \int_a^b q(x) \cdot P_{n+1}(x) \cdot w(x)\, dx + \int_a^b r(x) \cdot w(x)\, dx \tag{3.17}$$

l'integrale contenente $q(x)$ e' nullo perche $P_{n+1}(x)$ e' ortogonale a qualsiasi polinomio di grado inferiore ad $n+1$.

Utilizzando per $r(x)$ lo sviluppo in polinomi di $Lagrange$ (vedi appendice):

$$r(x) = \sum_{i=1}^{n+1} r(x_i) l_i(x) = \sum_{i=1}^{n+1} p(x_i) l_i(x) \tag{3.18}$$

e sostituendolo in (3.17):

$$\int_a^b p(x) \cdot w(x)\, dx = \int_a^b r(x) \cdot w(x)\, dx = \sum_{i=1}^{n+1} w_i p(x_i) \quad , \quad w_i = \int_a^b l_i(x) \cdot w(x)\, dx \tag{3.19}$$

Le quantita' w_i sono costanti che non dipendono dalla funzione $f(x)$. Abbiamo dunque trovato una espressione del tipo (3.2) che e' esatta per qualsiasi polinomio di grado $2n+1$ (ovviamente sara' valida anche per polinomi di grado inferiore). In analogia con quanto visto per il metodo di $Simpson$ c'e' da aspettarsi, e non lo dimostreremo, che per una generica funzione $f(x)$ per cui valga lo sviluppo di Taylor l'errore sia proporzionale alla derivata di ordine $2n+2$ di $f(x)$:

$$\int_a^b f(x) \cdot w(x)\, dx \simeq \sum_{i=1}^{n+1} w_i f(x_i) \tag{3.20}$$

$$e = \alpha \cdot \left(\frac{b-a}{2}\right)^{2n+3} \cdot f^{(2n+2)}(\xi) \cdot \frac{1}{(2n+2)!} \quad , \quad \xi \in [a,b] \tag{3.21}$$

dove α e' una costante indipendente da $f(x)$ e da n.

Per confrontare il metodo di $Gauss$ con quello di $Simpson$ ci si deve mettere in condizioni di parita' nel numero di calcoli da effettuare; questo vuol dire che si deve far coincidere l'intervallo $[a,b]$ con l'intervallo $[x_i - h, x_i + h]$ e si deve porre $n = 2$; in questo caso l'errore sara' dell'ordine h^7, contro un errore dell'ordine h^5 nel caso di $Simpson$. Per valori di n maggiori il confronto sara' ancor piu' favorevole per il metodo di $Gauss$.

Quando si vuole applicare il metodo di $Gauss$ al calcolo dell'integrale (3.1) si possono presentare tre casi:

- l'intervallo $[a, b]$ e' finito; in tal caso si possono utilizzare i polinomi di $Legendre$ che sono ortogonali sull'intervallo $[-1, 1]$ con una funzione peso $w(x) = 1$; ovviamente bisogna preventivamente effettuare un cambiamento di variabile che trasformi l'intervallo $[a, b]$ nell'intervallo $[-1, 1]$.

- l'intervallo $[a, b]$ e' semi-infinito; in tal caso possono essere utilizzati i polinomi di $Laguerre$ che sono ortogonali sull'intervallo $[0, +\infty]$ con una funzione peso $w(x) = exp(-x)$; in questo caso, oltre alla trasformazione dell'intervallo di integrazione, bisogna effettuare una trasformazione sulla funzione integranda:

$$\int_a^b f(x)\,dx = \int_a^b g(x) \cdot exp(-x)\,dx \quad , \quad g(x) = f(x) \cdot exp(x). \tag{3.22}$$

- l'intervallo $[a, b]$ e' infinito; i polinomi da utilizzare sono quelli di $Hermite$, ortogonali sull'intervallo $[-\infty, +\infty]$ secondo la funzione peso $exp(-x^2)$.

in tutti e tre i casi i punti x_i e i pesi w_i possono essere calcolati una volta per tutte; i valori numerici possono essere trovati in vari testi. Esistono anche, per punti e pesi gaussiani, delle formule di ricorrenza analitiche.

3.4 Metodo Monte Carlo

Un modo totalmente diverso, rispetto ai metodi analitici visti finora, di risolvere il problema dell'integrazione approssimata utilizza le proprieta' statistiche delle medie di variabili casuali e il $Teorema\ del\ limite\ centrale$. Il metodo si applica indifferentemente ad integrali in una o in piu' variabili e in questo paragrafo, e solo in questo, per non appesantire la notazione, intenderemo con x un punto di uno spazio di qualsiasi dimensione e gli integrali fatti su x come integrali multipli.

- *Teorema del limite centrale*

  Consideriamo una variabile casuale x, che abbia media μ_x e varianza σ_x^2 finite, e sia $g(x)$ la sua funzione di distribuzione. Allora la media fatta su un campione di N elementi estratti dalla distribuzione definita da $g(x)$:

$$\overline{x} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i \tag{3.23}$$

  e', nel limite $N \to \infty$, distribuita gaussianamente con media μ_x e varianza $\frac{\sigma_x^2}{N}$.

Il teorema va inteso in questo senso: consideriamo la media di N valori di una variabile casuale distribuita secondo la $g(x)$; e consideriamo tante successive estrazioni di N valori ciascuna e le relative medie. Allora la distribuzione di queste medie sara' una gaussiana coi parametri sopra indicati.

Esso si applica anche a una funzione di variabile casuale. Sia $f(x)$ tale funzione; il teorema afferma allora che la quantita':

$$\overline{f} = \frac{1}{N} \cdot \sum_{i=1}^{N} f(x_i) \tag{3.24}$$

e' distribuita gaussianamente attorno alla media:

$$\mu_f = \int f(x) \cdot g(x)\, dx \tag{3.25}$$

con varianza:

$$\sigma_{\overline{f}}^2 = \frac{\sigma_f^2}{N} \quad , \quad \sigma_f^2 = \int (f(x) - \mu_f)^2 \cdot g(x)\, dx \tag{3.26}$$

(sempre purche' gli integrali in (3.25) e (3.26) esistano).

Se poi x e' distribuita uniformemente entro un volume V si avra': $g(x) = \frac{1}{V}$, e quindi:

$$\mu_f = \frac{1}{V} \cdot \int f(x)\, dx \quad , \quad \sigma_f^2 = \frac{1}{V} \cdot \int (f(x) - \mu_f)^2\, dx \tag{3.27}$$

Se ora si ha a disposizione una successione di valori casuali estratti da una distribuzione uniforme entro il volume V, potremo calcolare $\overline{f}$ e scrivere:

$$\overline{f} = \frac{1}{N} \cdot \sum_{i=1}^{N} f(x_i) \simeq \mu_f = \frac{1}{V} \cdot \int f(x)\, dx \tag{3.28}$$

dove l'uguaglianza approssimata va intesa nel senso statistico che abbiamo discusso prima. Quindi scriveremo:

$$\int f(x)\, dx \simeq \frac{V}{N} \cdot \sum_{i=1}^{N} f(x_i) = I \tag{3.29}$$

ed assegneremo a questa stima dell'integrale un errore, statistico, dato dalla larghezza della distribuzione gaussiana:

$$e = \sigma_I = V \cdot \sigma_{\overline{f}} = V \cdot \frac{\sigma_f}{\sqrt{N}} \tag{3.30}$$

In questa espressione dell'errore il termine σ_f puo' anche essere molto grande. Esistono tuttavia delle tecniche per ridurlo che qui non descriveremo (il modo piu' semplice e' quello di suddividere il volume d'integrazione in molti sottovolumi; in ciascun sottovolume σ_f sara' piu' piccolo, con un guadagno netto sull'errore totale).

Ci vogliamo invece soffermare sull'andamento dell'errore al variare di N; il metodo $Monte\ Carlo$ puo' sembrare un modo veramente rozzo di calcolare integrali, e questo e' confermato dalla convergenza con $\frac{1}{\sqrt{N}}$, che e' molto lenta. Tuttavia questo andamento e' lo stesso qualunque sia la dimensione dello spazio. Per i metodi analitici descritti in precedenza, invece, l'andamento con N varia al variare della dimensione dello spazio; per rendere esplicita tale dipendenza, e per semplicita' di calcolo, considereremo il metodo del $punto\ di\ mezzo$, ma le stesse considerazioni valgono per tutti gli altri. L'errore e' calcolato nelle appendici; sostituendo:

$$h = \left(\frac{V}{N}\right)^{\frac{1}{d}} \tag{3.31}$$

nella (3.44) otteniamo per l'errore del metodo del punto di mezzo (d indica la dimensione dello spazio):

$$e_{pm} = \frac{1}{24} \cdot V^{1+\frac{2}{d}} \cdot \frac{1}{N^{\frac{2}{d}}} \cdot \delta \tag{3.32}$$

Le due espressioni (3.30) e (3.32) sono scritte in una forma appropriata per il confronto: σ_f e δ non dipendono da V ne' da N; l'andamento con V e', per d abbastanza grande, all'incirca lo stesso.

Ma per $d > 4$ l'andamento con N e' piu' favorevole per il metodo statistico. Questo valore di 'soglia' per d dipende dal metodo analitico usato; ad esempio per il metodo di $Simpson$ sara' 8 e ancora maggiore sara' per il metodo di $Gauss$, ma esistera' per qualunque metodo. Arriviamo dunque alla conclusione che il metodo $Monte\ Carlo$ e' competitivo con gli altri metodi per la stima di integrali multidimensionali. Lasciamo al lettore la ricerca di una spiegazione qualitativa del fenomeno.

Lasciamo infine al lettore il seguente esercizio:

- Consideriamo una serie di linee in un piano, tutte parallele tra loro ed equidistanti; sia D la distanza tra due linee succesive. E l'evento casuale costituito dal lancio di un bastoncino di lunghezza D da una grande altezza su questo piano. Dimostrare che la probabilita' che il bastoncino tocchi una delle linee e' $\frac{2}{\pi}$.

Si tratta di un metodo puramente statistico di 'misurare' π, che fu scoperto ed attuato alla fine del diciottesimo secolo ($ago\ di\ Buffon$); calcolare il numero di lanci necessario per avere la terza cifra decimale di π.

3.5 Appendici

3.5.1 Polinomi ortogonali

Data una funzione peso $w(x)$ non negativa nell'intervallo $[a, b]$ esiste un sistema completo di polinomi $P_i(x)$ ortogonali fra loro secondo il prodotto scalare definito, per due funzioni $f(x)$ e $g(x)$, da:

$$< f, g >= \int_a^b f(x) \cdot g(x) \cdot w(x)\, dx \tag{3.33}$$

$$< P_i, P_j >= \delta_{ij} \tag{3.34}$$

dove δ_{ij} e' il simbolo di Kronecker; a e b possono anche essere infiniti. Ovviamente $w(x)$ deve soddisfare condizioni tali che l'integrale esista per ogni i e j.

Diamo, senza dimostrarle, le tre proprieta' dei polinomi ortogonali che abbiamo usato nel testo:

- ogni polinomio di grado k puo' essere scritto come combinazione lineare di $P_0, P_1, P_2, ..., P_k$ (completezza).
- $P_i(x)$ e' ortogonale a $qualunque$ polinomio di grado minore di i:

$$< P_i, q_j >= 0 \quad , \quad j < i \tag{3.35}$$

- gli zeri di $P_i(x)$ sono tutti semplici ed interni ad $[a, b]$.

3.5.2 Formula di interpolazione di Lagrange

Dato un insieme di $n + 1$ punti distinti x_i, i polinomi di $Lagrange\ l_k(x)$ sono definiti da:

$$l_k(x) = \prod_{\substack{i=1 \\ i \neq k}}^{n+1} \frac{x - x_i}{x_k - x_i} \quad , \quad k = 1, 2, ..., n + 1. \tag{3.36}$$

Si ha evidentemente $l_k(x_j) = \delta_{kj}$; utilizzando questa proprieta' si puo' dimostrare facilmente che un generico polinomio di grado al piu' n, $p_j(x)$, puo' essere scritto nella forma:

$$p_j(x) = \sum_{i=1}^{n+1} p_j(x_i) \cdot l_i(x) \quad , \quad j \leq n. \tag{3.37}$$

infatti da entrambi i lati dell'uguaglianza abbiamo due polinomi di grado al piu' n che coincidono in $n+1$ punti distinti. Per il principio di identita' dei polinomi essi devono essere identici.

I polinomi di $Lagrange$ possono essere utilizzati per scrivere il polinomio di grado n che interpola una data funzione $f(x)$ in $n+1$ punti; esso sara' dato da:

$$\sum_{i=1}^{n+1} f(x_i) \cdot l_i(x) \tag{3.38}$$

3.5.3 Metodo del punto di mezzo in piu' dimensioni

L'estensione a piu' dimensioni delle formule di integrazione approssimata e' immediata nel metodo; tuttavia si presenta un problema che e' assente nel caso unidimensionale: mentre l'intervallo unidimensionale e' suddivisibile esattamente in un numero intero di intervalli elementari, in piu' dimensioni la suddivisione di un generico volume in parallelepipedi elementari non e' possibile. Poiche' siamo solo interessati a ricavare l'espressione dell'errore, non discuteremo qui il problema, ma considereremo come cella elementare un cubo e come volume di integrazione un parallelepipedo suddivisibile esattamente in un numero intero di celle elementari.

Il metodo del $punto\ di\ mezzo$ consiste nell'approssimare la funzione in ciascuno cubo elementare col valore assunto nel suo centro.

Indichiamo con d la dimensione dello spazio e con $\underline{x}$ un punto di tale spazio:

$$\underline{x} \equiv (x_1, x_2, x_3, ..., x_d) \tag{3.39}$$

Indichiamo anche con:

$$\underline{x}_0 \equiv (x_{0,1}, x_{0,2}, x_{0,3}, ..., x_{0,d}) \tag{3.40}$$

il centro di una generica cella elementare, e con h il suo lato.

Lo sviluppo di Taylor in un punto di tale cella da':

$$f(\underline{x}) = f(\underline{x}_0) + \sum_{i=1}^{d} \frac{\partial f(\underline{x}_0)}{\partial x_i} \cdot (x_i - x_{0,i}) + \frac{1}{2} \cdot \sum_{i,j=1}^{d} \frac{\partial^2 f(\underline{\xi}_0)}{\partial x_i \partial x_j} \cdot (x_i - x_{0,i}) \cdot (x_j - x_{0,j}) \tag{3.41}$$

l'integrale sulla cella elementare del primo termine da' $f(\underline{x}_0) \cdot h^d$; quello del secondo e' nullo (l'integrale su ogni variabile x_i va fatto da $x_{0,i} - \frac{h}{2}$ a $x_{0,i} + \frac{h}{2}$); quello del terzo da' l'errore: gli integrali dei termini con $i \neq j$ sono nulli, i restanti termini danno:

$$e_0 = \frac{1}{2} \cdot \sum_{i=1}^{d} \frac{\partial^2 f(\underline{\xi}_0)}{\partial x_i^2} \cdot \frac{1}{12} \cdot h^3 \cdot h^{d-1} = \frac{1}{24} \cdot h^{d+2} \cdot \sum_{i=1}^{d} \frac{\partial^2 f(\underline{\xi}_0)}{\partial x_i^2} \tag{3.42}$$

Per ottenere l'errore globale dobbiamo sommare su tutte le celle che compongono il volume d'integrazione; sia V tale volume ed N il numero totale di punti:

$$e = \frac{1}{24} \cdot h^{d+2} \cdot \sum_{j=1}^{N} \sum_{i=1}^{d} \frac{\partial^2 f(\underline{\xi}_j)}{\partial x_i^2} = \frac{1}{24} \cdot h^{d+2} \cdot N \cdot \sum_{i=1}^{d} \frac{\partial^2 f(\underline{\xi})}{\partial x_i^2} \tag{3.43}$$

dove $\underline{\xi}$ e' un punto appartenente al volume d'integrazione. Indicando con δ la sommatoria in (3.43) e sostituendo $N = \frac{V}{h^d}$:

$$e = \frac{1}{24} \cdot h^2 \cdot V \cdot \delta \tag{3.44}$$

Queste due ultime espressioni rendono esplicita la dipendenza dal numero di punti utilizzati o dal volume di integrazione (δ non dipende da queste due quantita'); sono simili all'espressione unidimensionale, con V sostituito dall'ampiezza dell'intervallo d'integrazione.

Capitolo 4

IL METODO MONTE CARLO

GENERAZIONE DI VARIABILI PSEUDOCASUALI

4.1 Simulazione di eventi casuali

I risultati di misure fisiche sono descritti da variabili casuali di cui, almeno nei casi più semplici, siamo in grado di calcolare le funzioni di distribuzione. Ad esempio per la diffusione coulombiana di un elettrone da parte di un atomo, la densità di probabilità dell'angolo di diffusione si può calcolare a partire dalla sezione d'urto di Rutherford.

Consideriamo un fascio di particelle monoenergetiche che attraversano un certo spessore di un materiale omogeneo e subiscono, per esemplificare, solo diffusioni elastiche a corto range. All'uscita dal materiale la distribuzione in posizione, energia e angolo delle particelle dipenderà dallo spessore e dalla densità del materiale. Il calcolo del valor medio di queste quantità è abbastanza agevole; lo è meno quello delle distribuzioni intorno a questi valori medi. Possiamo pensare che, se il numero medio di singole interazioni con gli atomi del materiale è grande, valga il teorema del limite centrale e tutte le quantità siano distribuite gaussianamente intorno ai loro valori medi. Ma, come abbiamo già detto, il teorema fallisce nel predire le code estreme delle distribuzioni, cioè le grandi fluttazioni rispetto ai valori medi.

Ancor più complesso è il caso in cui il numero medio di interazioni è piccolo o lo spessore attraversato è riempito da diversi materiali con una geometria complicata.

Problemi di questo tipo possono essere risolti numericamente mediante una $simulazione$ al calcolatore della successioni dei processi elementari (le singole interazioni). Supponiamo di avere a disposizione un algoritmo che ci permette di estrarre una successione di valori numerici casuali e distribuiti secondo una distribuzione assegnata. Utilizzando questo algoritmo potremo $simulare$ il percorso di una singola particella all'interno del materiale: noto il libero cammino medio, estraendo una variabile casuale distribuita esponenzialmente stabiliremo il punto in cui essa subirà un diffusione, dopo di che stabiliremo l'angolo di diffusione generando una variabile distribuita secondo la sezione d'urto differenziale, e così via fino all'uscita dal materiale.

Ripetendo lo stesso processo per molte particelle otterremo una distribuzione in frequenze delle quantità all'uscita dal materiale che approssimerà la densità di probabilità tanto meglio quanto maggiore è il numero di particelle considerate.

Altri esempi di eventi casuali che possono essere simulati al calcolatore sono l'estrazione dei numeri di una $roulette$ (da cui il nome del metodo Monte Carlo) e il traffico automobilistico di una città : in questo caso una simulazione al calcolatore che parta da flussi di $input$ noti permette di studiare gli effetti dei divieti di circolazione, della presenza dei semafori, e di altre variabili su cui sia

possibile intervenire.

Ovviamente i problemi discussi in precedenza possono anche essere risolti analiticamente e ricondotti al calcolo di integrali, tanto più complessi quanto più complesso è il problema in esame. Il metodo che abbiamo descritto si riconduce quindi al calcolo di un integrale col metodo Monte Carlo.

Nei paragrafri seguenti discuteremo gli algoritmi numerici per la generazione di variabili $pseudo$-$casuali$.

4.2 Generazione di variabili pseudocasuali

Per $generare\ una\ variabile\ casuale$ intendiamo trovare una successione di valori $z_1, z_2, z_3, ..., z_n, ...$ che siano $casuali$ nel senso che valori successivi non sono correlati ai precedenti, e quindi non possono essere calcolati a partire da essi; e che seguano una data legge di distribuzione, nel senso che, per n grandi, la distribuzione delle frequenze degli z_i segue, entro gli errori, una legge assegnata $f(z)$.

Ora il calcolatore, essendo ripetitivo, e' lo strumento meno adatto per questo scopo. Tutto cio' che si puo' fare e' trovare un algoritmo per generare, a partire da un valore iniziale, una successione in cui le correlazioni tra valori successivi sono cosi' tenui che, per gli scopi del problema che si vuole risolvere, siano trascurabili. Una successione di questo tipo e' $ripetitiva$, nel senso che, partendo da uno stesso valore iniziale si riottiene la stessa successione; e $periodica$, nel senso che esiste comunque un m tale che $z_{m+1} = z_1$ e, quindi, a partire da z_{m+1} la succesione si ripete. La periodicita' e' legata al fatto che per il calcolo viene utilizzato un numero finito di bit; ne vedremo un esempio in seguito.

In questo senso i valori generati con un algoritmo di questo tipo sono detti piu' propriamente $pseudocasuali$.

Dal punto di vista del programmatore la ripetitivita' e' un vantaggio, perche' consente controlli e confronti in fase di sviluppo dei programmi. Per quanto riguarda la periodicita' bisogna che l'applicazione per la quale si vuole utilizzare l'algoritmo richieda un numero di valori molto inferiore al periodo.

Nel paragrafo successivo discuteremo l'algoritmo piu' semplice per la generazione di variabili distribuite uniformemente; questo, come vedremo nei paragrafi seguenti, costituisce la base per la generazione di variabili distribuite secondo una funzione qualsiasi.

4.3 Generazione di variabili pseudocasuali distribuite uniformemente tra 0 e 1.

Alla base del metodo che discuteremo sta l'idea intuitiva che il valore del resto di una divisione e' poco 'correlato' coi valori del dividendo e del divisore. Consideriamo un valore iniziale intero I_0 (il $seme$ del generatore) ed applichiamo la relazione di ricorrenza :

$$I_i = a \cdot I_{i-1} + b \qquad (mod\ \ m) \tag{4.1}$$

dove a , b, m sono costanti fissate intere e con $(mod\ m)$ intendiamo il resto della divisione della quantita' a destra per m. Tutti gli I_i così definiti sono interi compresi tra 0 ed $m-1$; i numeri:

$$z_i = \frac{I_i}{m-1} \tag{4.2}$$

saranno compresi tra 0 e 1.

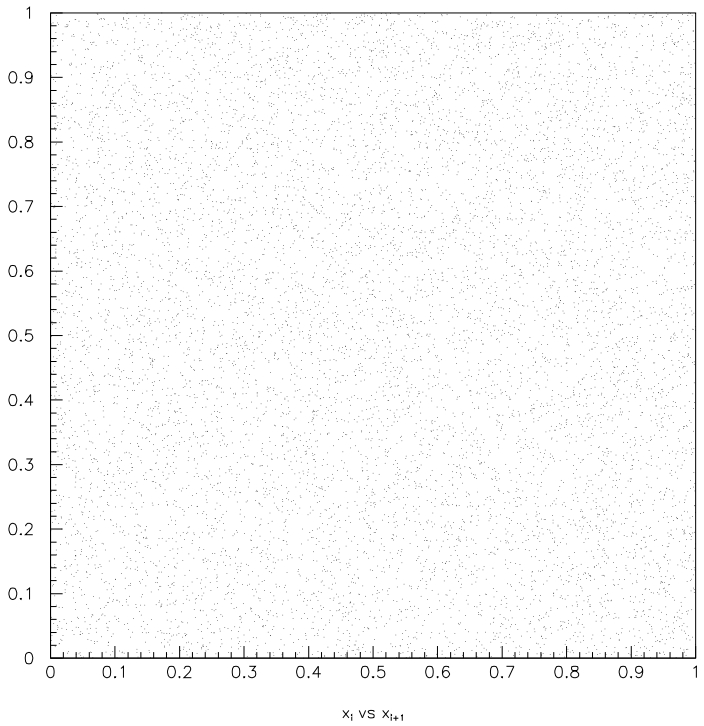


Figura 4.1: Distribuzione di x_i in funzione di x_{i+1} per 10000 termini della succesione di numeri pseudocasuali ottenuta con un generatore congruente .

E' evidente che questa succesione deve essere periodica, perche' puo' fornire al piu' m valori distinti; il periodo sara' anzi inferiore ad m, a seconda dei valori di a , b , I_0.

Generalmente, quando si utilizza un calcolatore a t bit si sceglie $m = 2^t$, perche' cio' rende particolarmente semplice il calcolo del resto della divisione: basta eseguire l'operazione $a \cdot I_{i-1} + b$ in un registro a t bit; tale registro conterra' direttamente il resto della divisione.

La $pseudocasualita'$ della successione degli z_i sta nel fatto che essi riempiranno uniformemente l'intervallo $[0, 1]$ e che la posizione di z_{i-1} in questo intervallo ha poca o nessuna correlazione con quella di z_i. In altre parole, per un insieme di z_{i-1} raggruppati in un piccolo sottointervallo di $[0, 1]$, gli z_i si distribuiranno uniformemente sull'intero intervallo (vedi fig. 4.1) [1].

Queste proprieta' vanno ovviamente verificate e dipendono certamente da a, b , m e dal valore iniziale scelto; non discuteremo questi temi.

Generalmente le librerie dei calcolatori comprendono una funzione che contiene dei valori opportuni di a , b e m; tale funzione va utilizzata assegnando il valore iniziale secondo regole indicate dal costruttore.

Il metodo definito da (4.1) viene detto $congruente$ nel caso $b = 0$ e $congruente\ misto$ se $b \neq 0$. Sono stati sviluppati vari altri metodi che richiedono operazioni piu' elaborate, ma tutti sono basati sul del calcolo di resti di divisioni e tutti sono $pseudocasuali$ nel senso detto prima.

[1] Ovviamente lo strumento matematico per rendere quantitative queste considerazioni è la *funzione di correlazione*.

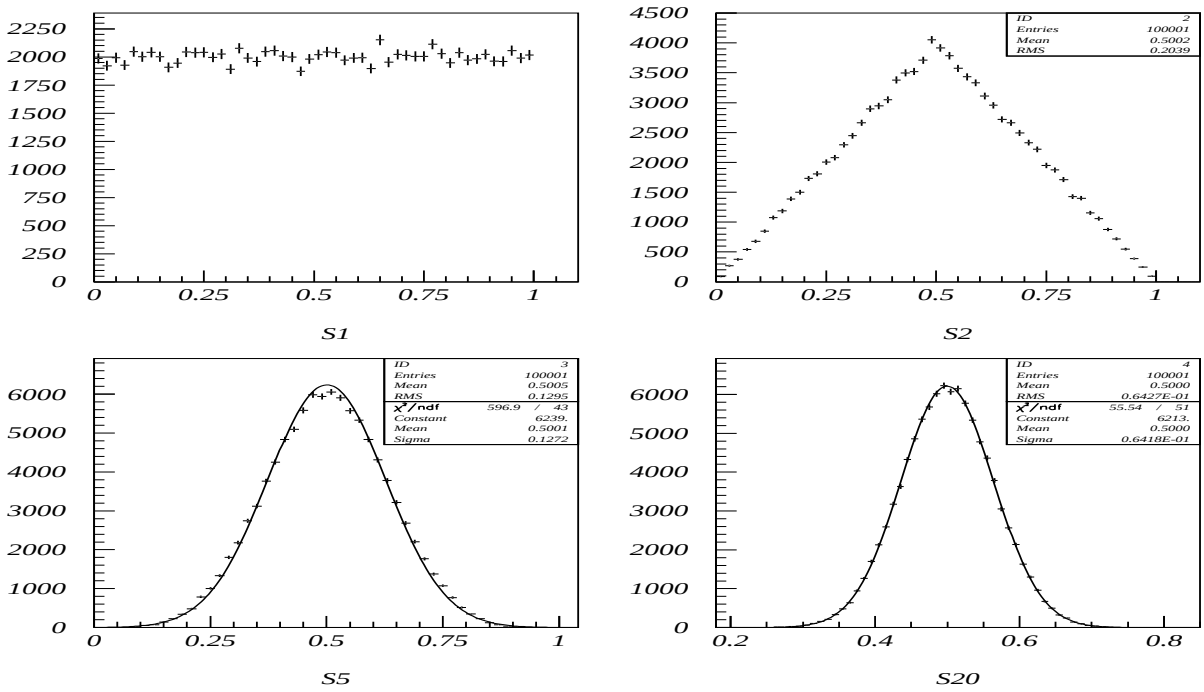


Figura 4.2: Il *Teorema del limite centrale* all'opera: istogramma delle frequenze di una variabile pseudocasuale distribuita uniformemente tra 0 e 1 generata col metodo *congruente* ($S1$) e della media di $2, 5, 20$ di queste variabili ($S2, S5, S20$ rispettivamente). Le curve sovrapposte rappresentano il risultato del fit gaussiano.
Si lascia come esercizio la verifica che il valore della varianza delle distribuzioni gaussiane è in accordo con la previsione del teorema e la dimostrazione che la media di due variabili casuali distribuite uniformemente è distribuita secondo una funzione triangolare.

4.4 Generazione di variabili pseudocasuali secondo una distribuzione assegnata.

A partire da un generatore di una variabile pseudocasuale distribuita uniformemente tra 0 e 1 e' possibile ottenere un generatore di una variabile pseudocasuale distribuita secondo una distribuzione assegnata. In questo paragrafo discuteremo alcuni dei metodi piu' semplici. Indicheremo con z la variabile distribuita uniformemente e con x una variabile pseudocasuale che varia nell'intervallo $[x_m, x_M]$ e la cui funzione di distribuzione e' $f(x)$.

Osserviamo innanzitutto che la variabile x ottenuta mediante la trasformazione:

$$x = x_m + z \cdot (x_M - x_m) \tag{4.3}$$

e' distribuita uniformemente tra x_m e x_M.

4.4.1 Metodo HIT OR MISS

Consideriamo una distribuzione bidimensionale uniforme nel rettangolo indicato in figura 4.3.

E' possibile ottenere un generatore di valori (x, y) pseudocasuali distribuiti uniformemente a partire da quello di due variabili z e t distribuite uniformemente nell'intervalllo $[0, 1]$ mediante la

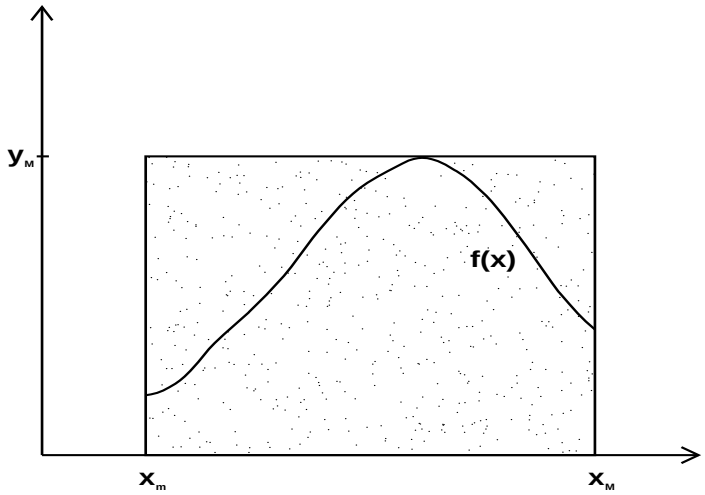


Figura 4.3:

trasformazione:

$$x = x_m + z \cdot (x_M - x_m) \tag{4.4}$$

$$y = t \cdot y_M \tag{4.5}$$

(data la sequenza z_1 ,z_2 ,z_3 ,z_4 , ... si utilizzano questi valori per calcolare alternativamente x_1 ,y_1 ,x_2 ,y_2 , ... , mediante (4.4) e (4.5).)

Se ora di tutti i punti (x_i, y_i) cosi' generati utilizziamo solo quelli per cui $y_i \leq f(x_i)$ (cioe' accettiamo solo i punti che in figura stanno al di sotto della curva che rappresenta $f(x)$) le x_i di tali punti saranno distribuite secondo la funzione di distribuzione $f(x)$.

L'algoritmo da utilizzare e' il seguente:

[1] ESTRAZIONE DI z_i
[2] ESTRAZIONE DI z_{i+1}
[3] SE $z_{i+1} \cdot y_M \geq f(z_i)$ TORNA A [1]
[4] $x_j = x_m + z_i \cdot (x_M - x_m)$

L' applicazione ripetuta di questo algoritmo produrra' una successione di valori x_j distribuiti secondo la $f(x)$.

Notiamo che questo metodo permette di ottenere un valore approssimato per l' integrale di $f(x)$:

$$\int_{x_m}^{x_M} f(x)\, dx \simeq \frac{N_a}{N_g} \cdot S \tag{4.6}$$

dove N_a e' il numero di punti accettati, N_g quello di punti generati ed S e' la superficie del rettangolo in figura. L' errore sul valore approssimato sara' $\frac{\sqrt{N_a}}{N_g} \cdot S$.

4.4.2 Generazione a partire da una primitiva di $f(x)$

Il metodo si basa sul fatto che, data una funzione di distribuzione $f(x)$ normalizzata, la variabile casuale $z(x)$ definita da:

$$z(x) = \int_{x_m}^{x} f(t)\, dt \tag{4.7}$$

e' distribuita uniformemente tra 0 e 1; detta infatti $g(z)$ la funzione di distribuzione della variabile z, si ha:

$$g(z) = f(x) \left| \frac{dx}{dz} \right| = 1 \tag{4.8}$$

d' altra parte, per x che varia tra x_m e x_M, z assume tutti i valori compresi tra 0 e 1.

Se indichiamo ora con $F(x)$ una primitiva di $f(x)$, si ha:

$$z(x) = F(x) - F(x_m) \tag{4.9}$$

e

$$x = F^{-1}(z + F(x_m)). \tag{4.10}$$

Se si sceglie la primitiva F in modo tale che $F(x_m) = 0$:

$$x = F^{-1}(z) \tag{4.11}$$

la (4.10) o la (4.11) costituiscono dunque una trasformazione tra una variabile distribuita uniformemente z e una variabile x distribuita secondo la funzione $f(x)$.

Il metodo e' dunque applicabile quando F e' nota analiticamente o calcolabile in modo semplice (ad esempio calcolo numerico dell' integrale (4.7) o interpolazione a partire da una tabulazione dei valori di $F(x)$).

Ad esempio, la variabile x definita dalla trasformazione:

$$x = -\ln z \tag{4.12}$$

è distribuita esponenzialmente.

4.4.3 Generatori gaussiani

- Se z e t sono due variabili casuali indipendenti e distribuite uniformemente tra 0 e 1, le due variabili casuali x e y definite dalla trasformazione:

  $$x = \sqrt{-2 \cdot \ln z} \cdot \cos(2\pi \cdot t) \tag{4.13}$$

  $$y = \sqrt{-2 \cdot \ln z} \cdot \sin(2\pi \cdot t) \tag{4.14}$$

  sono indipendenti e distribuite normalmente con media 0 e deviazione standard 1 [esercizio!].

- Secondo il *Teorema del limite centrale*, la media di n variabili distribuite umiformemente tra 0 e 1 e', per $n \to \infty$, distribuita gaussianamente tra 0 e 1 con media $\frac{1}{2}$ e deviazione standard $\frac{\sigma_0}{\sqrt{n}}$; $\sigma_0 = \frac{1}{\sqrt{12}}$ e' la deviazione standard della distribuzione uniforme. Una buona approssimazione della distribuzione gaussiana si ottiene gia' per $n \sim 10$.

  Questo generatore ha il difetto di non poter riprodurre le code estreme della distribuzione.

Una semplice trasformazione su una variabile gaussiana permette poi di ottenere un' altra variabile gaussiana di media e deviazione standard qualsiasi.

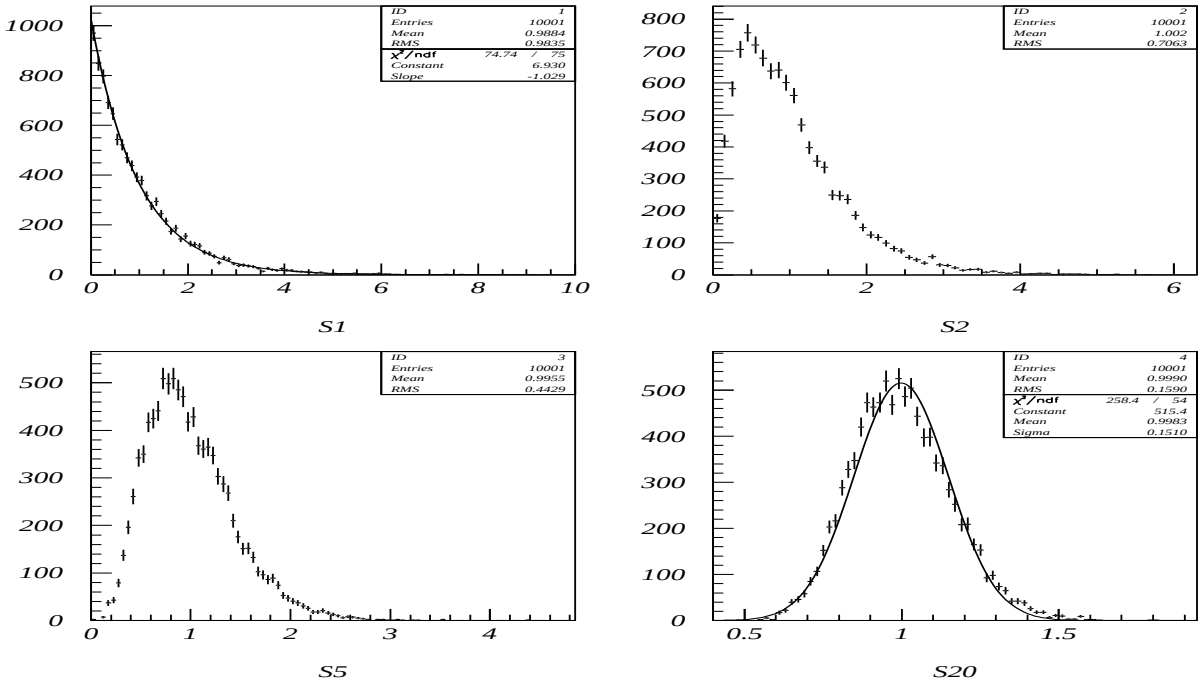


Figura 4.4: Come la figura (4.3) , ma per una variabile distribuita esponenzialmente generata mediante la trasformazione (4.12).

4.4.4 Un semplice esempio di simulazione : L'*Ago di Buffon*

Alla fine del diciottesimo secolo il matematico francese Buffon ideò un metodo puramente statistico per calcolare π. La descrizione del sistema viene data nell'esercizio seguente:

- Consideriamo una serie di linee in un piano, tutte parallele tra loro ed equidistanti; sia D la distanza tra due linee succesive. E l'evento casuale costituito dal lancio di un bastoncino di lunghezza D da una grande altezza su questo piano. Dimostrare che la probabilita' che il bastoncino tocchi una delle linee e' $\frac{2}{\pi}$. Calcolare il numero di lanci necessari per avere la terza cifra decimale di π.

Il calcolo di questa probabilità si riduce a quello di un'integrale, quindi il metodo è di fatto una semplice applicazione del metodo *Monte Carlo* per il calcolo di integrali. Il sistema può essere facilmente simulato al calcolatore.

Nel programma seguente viene simulato l'esperimento di Buffon. **ran(iseed)** è la funzione di sistema che esegue l'estrazione col metodo *congruente* e ritorna un valore numerico compreso tra 0 e 1. La variabile **iseed** contiene inizialmente il seme del generatore, scelto dall'utente, e viene aggiornata dalla funzione **ran** ad ogni chiamata. **ran** viene utilizzata per generare, ad ogni estrazione , il punto d'impatto di un estremo del bastoncino e l'angolo rispetto ad un asse verticale.

```
implicit real*8 (a-h,o-z)

data pi/3.1415926536/
data iseed/5433789/
data d/1./

print *,'seme ? '
read *,iseed
print *,'numero di estrazioni ? '
```

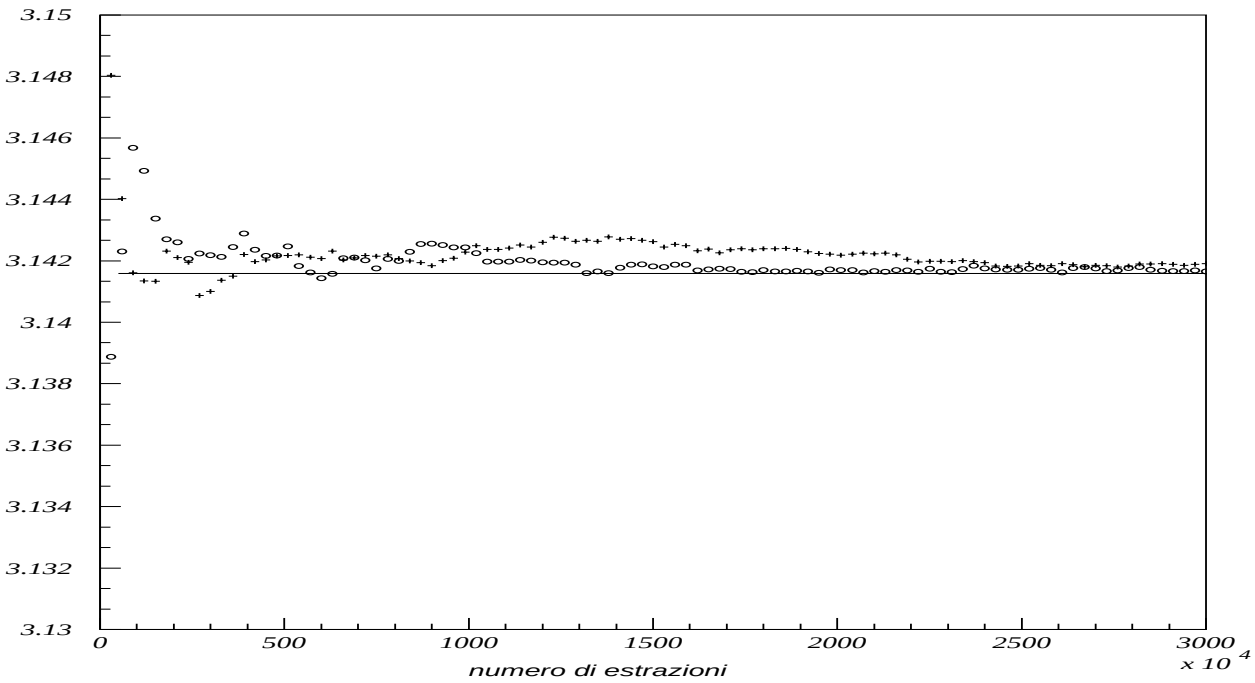


Figura 4.5: Il valore di π ottenuto col metodo dell'*Ago di Buffon* in funzione del numero di estrazioni per due diverse successioni di numeri casuali ottenute con due diversi semi.

```
read *,max

mfreq = max/100

 m = 0
 do j = 1 , max
  x1 = d * ran ( iseed )
  t  = pi * ran ( iseed )
  x2 = x1 + d * cos(t)
  if ( x2.ge.d .or. x2.le.0. ) m = m + 1
  if (mod(j,mfreq).eq.0) then
    pp = 2.d0*dfloat(j)/dfloat(m)
    epp = 2.d0*sqrt(dfloat(j))/dfloat(m)
    print * , j , pp,epp
    write(10,*) , j , pp,epp
  endif
 enddo

stop
end
```

4.5 Generazione di variabili casuali in più dimensioni - l'algoritmo di Metropolis

In una dimensione il problema della generazione secondo una qualsiasi funzione di distribuzione e' di semplice soluzione: nel peggiore dei casi e' sempre possibile utilizzare il metodo HIT OR MISS, eventualmente suddividendo l' intervallo delle x in modo da aumentare il rapporto tra il numero di valori accettati e quello di valori generati.

In piu' dimensioni questo rapporto puo' diventare molto piccolo, rendendo il metodo HIT OR MISS inefficiente in modo inaccettabile.

L'algoritmo di *Metropolis* risolve questo problema. Detto V il volume entro cui si vuole generare un insieme di punti $\mathbf{r_i}$ distribuiti secondo la funzione $f(\mathbf{r})$, l'algoritmo consiste nel generare inizialmente un insieme di $\mathbf{r_i}$ distribuiti uniformemente entro V (o meglio, se possibile, distribuiti secondo una funzione il più possibile vicina a $f(\mathbf{r})$). A questi punti vengono poi fatti eseguire un certo numero di passi casuali entro il volume V; lo spostamento in ciascun passo sarà dato da:

$$r'_{i,j} = r_{i,j} + \delta \cdot t_j \tag{4.15}$$

dove $r_{i,j}$ è la componente j-esima di $\mathbf{r_i}$ e t_j è una variabile casuale distribuita uniformemente tra -1 e 1; δ è una quantità fissata.

Definito questo passo dobbiamo introdurre una regola per accettarlo o respingerlo (cioè lasciare il punto nella posizione iniziale $\mathbf{r_i}$). Tale regola sarà scelta in modo che dopo un certo numero di passi si realizzi una condizione di equilibrio con una densità di punti data da $f(\mathbf{r})$. Possiamo immaginare questo insieme di punti come un fluido di densità $f(\mathbf{r})$ in movimento con una legge che lascia invariata questa densità.

Data una qualunque coppia di punti $\mathbf{r}$ ed $\mathbf{s}$ e detta $P(\mathbf{r},\mathbf{s})$ la densità di probabilità di effettuare uno spostamento da un intorno di $\mathbf{s}$ ad uno di $\mathbf{r}$, la condizione di equilibrio richiede che il numero di punti che vanno da $\mathbf{s}$ ad $\mathbf{r}$ sia uguale a quello di punti che vanno da $\mathbf{r}$ ad $\mathbf{s}$:

$$P(\mathbf{r},\mathbf{s}) \cdot f(\mathbf{s}) = P(\mathbf{s},\mathbf{r}) \cdot f(\mathbf{r}) \tag{4.16}$$

Ora $P(\mathbf{r},\mathbf{s})$ può essere scritta come il prodotto della densità di probabilità di *proporre* uno spostamento per quella di *accettare* tale spostamento; indicheremo quet'ultima con $A(\mathbf{r},\mathbf{s})$. La densità di probabilità di proporre lo spostamento da $\mathbf{s}$ ad $\mathbf{r}$ è uguale a quella per lo spostamento da $\mathbf{r}$ ad $\mathbf{s}$, perchè entrambe governate dalla (4.15); qundi la (4.16) si riduce a:

$$A(\mathbf{r},\mathbf{s}) \cdot f(\mathbf{s}) = A(\mathbf{s},\mathbf{r}) \cdot f(\mathbf{r}) \tag{4.17}$$

Definito il rapporto:

$$q(\mathbf{r},\mathbf{s}) = \frac{f(\mathbf{r})}{f(\mathbf{s})} \tag{4.18}$$

la (4.17) è soddisfatta dalla scelta:

$$A(\mathbf{r},\mathbf{s}) = \min(1, q(\mathbf{r},\mathbf{s})) \tag{4.19}$$

oppure:

$$A(\mathbf{r},\mathbf{s}) = \frac{q(\mathbf{r},\mathbf{s})}{1 + q(\mathbf{r},\mathbf{s})} \tag{4.20}$$

quindi,se utilizziamo la (4.19), accetteremo comunque il passo (4.15) se $f(\mathbf{r}') \geq f(\mathbf{r})$, mentre se $f(\mathbf{r}') < f(\mathbf{r})$ lo accetteremo con una probabilità data da q ($\mathbf{r}$',$\mathbf{r}$).

In quest'ultimo caso per decidere se accettare o no il passo, possiamo estrarre un numero casuale z distribuito uniformemente tra 0 e 1; se z risulterà minore di $q(\mathbf{r}',\mathbf{r})$ accetteremo il passo, in caso contrario lo respingeremo.

Nel programma seguente l'algoritmo di Metropolis viene applicato per generare in uno spazio bidimensionale 3000 punti distribuiti secondo la funzione:

$$\alpha \cdot \exp(-|x|) \cdot \exp(-|y|) \tag{4.21}$$

partendo da una distribuzione uniforme; l'algoritmo vero e proprio è codificato nella parte del programma indicata dai commenti, il resto serve per la gestione dell'output.

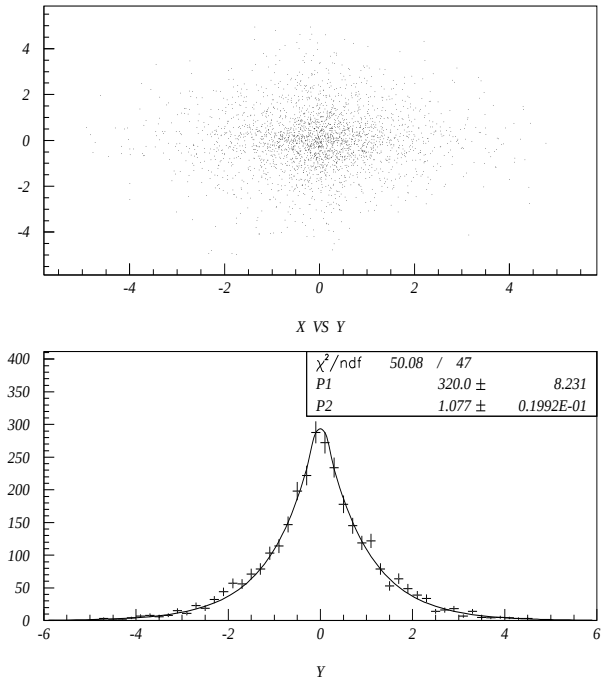


Figura 4.6: Distribuzione bidimensionale ottenuta con l'algoritmo di Metropolis.

nella figura 4.6 viene mostrato la distribuzione di punti che risulta scegliendo $\delta = 2$ e 150 passi per ogni punto.

Va notato che l'applicazione dell'algoritomo così come è stato descritto porta ad una distribuzione distorta, rispetto a quella richiesta, sui bordi di V, e questo effetto è tanto più grande quanto più è grande δ; questo viene dal fatto che non possiamo accettare passi che portino i punti fuori da V.

Nel caso in esame la scelta di un valore di δ piuttosto grande rispetto alla scala di variazione della funzione (4.21) risulta abbastanza efficace. In altri casi (ad esempio quando la distribuzione assume valori grandi sul bordo della regione di definizione) questa scelta non è possibile; il lettore è invitato ad immaginare le possibili soluzioni di questo problema.

```
      parameter(nh=20000)
      parameter(npoints=3000,xmax=5.,ymax=5.)
      character*15 names(2)
      common/pawc/h(nh)
      dimension x(npoints),y(npoints),g(npoints)

      data names / 'x ' , 'y ' /
      data iseed/1865433789/
      dimension var(2)
c
c               allocate memory for hbook
c

      call hlimit(nh)

      open(unit=29,file='metropolis.nt',access='direct',recl=1024,
     +     status='new')
c
c  inizializzazione n-upla

      call hbookn(100,'test',2,'//test',2000,names)
      call hrfile(29,'test','n')

c  ....   inizio algoritmo di Metropolis   .....
```

```
      print *,'passo max. e numero di passi:'
      read *,step,nstep
      do k = 1 , npoints
        x(k) = ( 2.*ran(iseed) - 1. ) * xmax
        y(k) = ( 2.*ran(iseed) - 1. ) * ymax
        g(k) = f(x(k),y(k))
      enddo

      do k = 1 , npoints
        do j = 1 , nstep
          xn = -10.*xmax
          yn = -10.*ymax
          do while ( xn.lt.-xmax .or.
     >               yn.lt.-ymax .or.
     >               xn.gt.xmax .or.
     >               yn.gt.ymax )
            xn = x(k) + step * ( ran(iseed) - .5 )
            yn = y(k) + step * ( ran(iseed) - .5 )
          enddo

          gg = f(xn,yn)
          rap = gg / g(k)
          sub = .false.
          if ( ran(iseed) .le. rap ) then
            x(k) = xn
            y(k) = yn
            g(k) = gg
          endif
        enddo      !       on nstep
        var(1) = x(k)
        var(2) = y(k)
        call hfn(100,var)
      enddo

c  ....   fine algoritmo di Metropolis   .....

      call hrout(0,icycle,' ')
      call hrend('test')

      stop
      end

      function f(x,y)
        f = exp(-abs(x))*exp(-abs(y))
      return
      end
```

Capitolo 5

METODI APPROSSIMATI PER LA SOLUZIONE DI EQUAZIONI DIFFERENZIALI ORDINARIE

5.1 Introduzione

Consideriamo una equazione differenziale ordinaria del primo ordine [1] :

$$y' = f(x, y) \tag{5.1}$$

con la condizione iniziale:

$$y(x_0) = y_0 \tag{5.2}$$

Il nostro scopo e' calcolare dei valori approssimati di $y(x)$ su una griglia di punti (normalmente equispaziati):

$$x_n = x_0 + n \cdot h \tag{5.3}$$

Indichiamo con y_n i valori approssimati di $y(x_n)$ ottenuti con qualche metodo di approssimazione. Tale metodo dovra' fornirci una relazione di ricorrenza tra gli y_n:

$$y_{n+1} = g(x_n, y_n) \tag{5.4}$$

che valga, approssimata ad un certo ordine in h, anche per $y(x_n)$:

$$y(x_{n+1}) = g(x_n, y(x_n)) + O(h^j) \quad [2] \tag{5.5}$$

$O(h^j)$ e' l'errore $locale$ dell'approssimazione, nel senso che, se y_n e' il valore esatto di $y(x_n)$, allora da (5.4) otterremo un valore per y_{n+1} che differisce da $y(x_{n+1})$ per un $O(h^j)$.

In realta' gli y_n che si ottengono reiterando piu' volte la (5.4) sono valori approssimati di $y(x_n)$; dopo un certo numero di passi avremo dunque un errore $globale$ che e' sempre maggiore di quello locale ed in generale, come vedremo, e' $O(h^{j-1})$. Il valore iniziale di y_0 nella (5.4) sara' fornito dalla condizione iniziale (5.2).

[1] Una equazione differenziale di ordine n può sempre essere ridotta ad un sistema di n equazioni differenziali del primo ordine ; le considerazioni che faremo nel seguito si applicano anche a questo sistema.

[2] indichiamo con $O(h^j)$ una quantita' che tende a zero come h^j.

5.2 Metodo di EULERO

Consideriamo lo sviluppo di Taylor:

$$y(x_{n+1}) = y(x_n) + h \cdot y'(x_n) + \frac{1}{2} \cdot h^2 \cdot y''(\xi) \quad , \quad \xi \in [x_n, x_{n+1}] \tag{5.6}$$

utilizzando la (5.1) otteniamo una relazione del tipo (5.5) (con $j = 2$):

$$y(x_{n+1}) = y(x_n) + h \cdot f(x_n, y(x_n)) + \frac{1}{2} \cdot h^2 \cdot y''(\xi) \tag{5.7}$$

la corrispondente relazione tra gli y_n sara':

$$y_{n+1} = y_n + h \cdot f(x_n, y_n) \tag{5.8}$$

l'errore locale e' dunque dell'ordine h^2.

5.3 Metodo di TAYLOR

Il metodo di $Eulero$ e' un caso particolare del metodo di $Taylor$; consideriamo lo sviluppo di Taylor fino ad un certo ordine k di derivazione e le espressioni di $y', y'', y'''...$ che ricaviamo dalla (5.1):

$$\begin{aligned} y'(x_n) &= f(x_n, y(x_n)) \\ y'' &= f_x + f_y \cdot y' = f_x + f_y \cdot f \\ y''' &= f_{xx} + 2 \cdot f_{xy} \cdot f + f_{yy} \cdot f^2 + f_x \cdot f_y + f_y^2 \cdot f \\ &\vdots \end{aligned} \tag{5.9}$$

dove coi pedici x e y indichiamo l'operazione di derivazione parziale. Sostituendo queste espressioni nello sviluppo di Taylor otterremo una generalizzazione delle (5.7) e (5.8) con un errore $O(h^k)$. Le espressioni risultano comunque piuttosto complicate; scopo degli altri metodi che descriveremo e' di ottenere la stessa precisione con calcoli piu' semplici rispetto al metodo di $Taylor$.

5.4 Metodo di RUNGE-KUTTA e altri metodi

Riscriviamo la (5.1) nella forma:

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x, y)\, dx \quad . \tag{5.10}$$

Il metodo di $Eulero$ corrisponde ad approssimare l'integrale in (5.10) con $h \cdot f(x, y(x_n))$, con un errore che abbiamo gia' calcolato. Se invece del punto estremo x_n utilizziamo il punto di mezzo (che indicheremo con $x_{n+\frac{1}{2}}$) dell'intervallo $[x_n, x_{n+1}]$ avremo un errore minore (esercizio!):

$$\int_{x_n}^{x_{n+1}} f(x, y)\, dx = h \cdot f(x_{n+\frac{1}{2}}, y(x_{n+\frac{1}{2}})) + O(h^3) \tag{5.11}$$

da cui:

$$y(x_{n+1}) = y(x_n) + h \cdot f(x_{n+\frac{1}{2}}, y(x_{n+\frac{1}{2}})) + O(h^3) \tag{5.12}$$

che non e' ancora una formula del tipo (5.5) perche' contiene il valore di y in un punto,$x_{n+\frac{1}{2}}$, che non e' della griglia. Possiamo pero' scrivere:

$$y(x_{n+\frac{1}{2}}) = y(x_n) + \frac{1}{2} \cdot h \cdot f(x_n, y(x_n)) + O(h^2) \tag{5.13}$$

sostituendo in (5.12) l'errore restera' $O(h^3)$ perche':

$$f(x_{n+\frac{1}{2}}, y(x_{n+\frac{1}{2}})) = f(x_{n+\frac{1}{2}}, y(x_n) + \frac{1}{2} h f(x_n, y(x_n)) + f_y \cdot O(h^2) \quad . \tag{5.14}$$

La relazione di ricorrenza che ne risulta:

$$y_{n+1} = y_n + h \cdot f(x_{n+\frac{1}{2}}, y_n + \frac{1}{2} \cdot h \cdot f(x_n, y_n)) \tag{5.15}$$

avra' quindi un errore $O(h^3)$. La (5.15) porta il nome di $metodo\ di\ Runge - Kutta\ del\ secondo\ ordine$ (l'errore globale e' $O(h^2)$). Altre espressioni ottenute in modo simile, del terzo e quarto ordine, possono essere trovate su vari testi.

Sui testi di analisi numerica possono anche essere trovati molti altri metodi di ordine superiore. Il modo di ottenerli non differisce di molto da quelli che abbiamo descritto, anche se le espressioni finali sono piu' complesse. Sono stati sviluppati in base al tipo di equazioni da risolvere e sono molto piu' efficienti del metodo di $Eulero$. In realta' quest'ultimo e' molto utile per descrivere in modo semplice le caratteristiche generali dei metodi di approssimazione, ma e' di un uso pratico molto limitato.

5.5 Instabilita' delle soluzioni numeriche - un esempio

In questo paragrafo considereremo il metodo di $Eulero$ applicato ad una semplice equazione risolvibile analiticamente e confronteremo la soluzione approssimata con quella esatta. Cio' allo scopo di evidenziare i problemi che si possono presentare e le strategie da seguire nell'applicare un metodo numerico. Considerazioni simili, anche se in forma matematica piu' complessa, possono essere fatti per tutti i metodi di approssimazione.

Consideriamo l'equazione:

$$y' = -\alpha \cdot y + \beta \quad , \quad y(0) = 0 \quad , \quad \alpha > 0 \tag{5.16}$$

il metodo di $Eulero$ ci da' la relazione di ricorrenza:

$$y_{n+1} = y_n + h \cdot (-\alpha \cdot y_n + \beta) \tag{5.17}$$

che riscriviamo nella forma:

$$y_{n+1} - y_n \cdot (1 - \alpha \cdot h) - h \cdot \beta = 0 \tag{5.18}$$

questa e' un'equazione alle differenze del primo ordine lineare (vedi appendice) la cui soluzione generale puo' essere scritta in forma analitica:

$$y_n = c \cdot (1 - \alpha \cdot h)^n + \frac{\beta}{\alpha} \tag{5.19}$$

dove c e' una costante arbitraria che va determinata a partire dalla condizione iniziale. Per la (5.16) si ha che $c = -\frac{\beta}{\alpha}$. Utilizzando il limite che definisce e (base dei logaritmi naturali) si verifica facilmente che la soluzione cosi' ottenuta tende , per $h \to 0$, alla soluzione esatta della (5.16).

Discuteremo ora una modifica al metodo di $Eulero$ che ha un errore inferiore rispetto ad esso, ma introduce soluzioni spurie, ed e' dunque inutilizzabile. Il metodo consiste nell'utilizzare l'approssimazione della differenza simmetrica per la derivata prima (il metodo di $Eulero$ consiste di fatto nell'utilizzare la differenza asimmetrica):

$$y'(x_n) = \frac{y(x_{n+1}) - y(x_{n-1})}{2 \cdot h} + O(h^2) \tag{5.20}$$

che conduce alla relazione:

$$y(x_{n+1}) = y(x_{n-1}) + 2 \cdot h \cdot y'(x_n) + O(h^3) \tag{5.21}$$

che e' analoga alla (5.7) e rispetto ad essa converge piu' rapidamente. Scriviamo ora la corrispondente equazione alle differenze per la (5.16):

$$\begin{aligned} y_{n+1} &= y_{n-1} + 2 \cdot h \cdot (-\alpha \cdot y_n + \beta) \\ y_{n+1} &+ 2 \cdot \alpha \cdot h \cdot y_n - y_{n-1} - 2 \cdot h \cdot \beta = 0 \end{aligned} \tag{5.22}$$

l'equazione caratteristica e le sue soluzioni sono date da:

$$\lambda^2 + 2 \cdot \alpha \cdot h \cdot \lambda - 1 = 0 \quad , \quad \lambda = -\alpha h \mp \sqrt{\alpha^2 h^2 + 1} \simeq -\alpha h \mp 1 \tag{5.23}$$

l'approssimazione e' valida per $h \ll \frac{1}{\alpha}$ e non modifica le conclusioni che stiamo per trarre.

La soluzione generale della (5.22) e' dunque data da:

$$\begin{aligned} y_n &= c_1 \cdot (1 - \alpha \cdot h)^n + c_2 \cdot (-1 - \alpha \cdot h)^n + \frac{\beta}{\alpha} \\ &= c_1 \cdot (1 - \alpha \cdot h)^n + c_2 \cdot (-1)^n \cdot (1 + \alpha \cdot h)^n + \frac{\beta}{\alpha} \end{aligned} \tag{5.24}$$

dove c_1 e c_2 sono due costanti arbitrarie da determinare a partire dalle condizioni iniziali. Come si vede abbiamo nuovamente un termine $c_1(1 - \alpha h)^n + \frac{\beta}{\alpha}$ che per $h \to 0$ tende all'esponenziale decrescente che e' soluzione esatta dell'equazione di partenza; ma ora si e' aggiunto un termine che invece tende ad un'esponenziale crescente moltiplicato per $(-1)^n$. questo termine cresce in valore assoluto all'aumentare di n e per di piu' cambia segno ad ogni passo; e' chiaramente un termine che non ha nulla a che fare con l'equazione originale ma e' stato introdotto dal metodo di approssimazione usato. Questo e' un esempio tipico di quella che viene detta $instabilita'$ del metodo numerico.

La stabilita' dei metodi di approssimazione andrebbe studiata caso per caso, sulla base delle equazioni a cui vanno applicati. In generale l'equazione alle differenze che ne risulta non e' risolubile analiticamente, per cui $convenzionalmente$ si studia la stabilita' nel caso della equazione $y' = \lambda x$ ed il metodo viene classificato in base ai valori di λ e di h per cui risulta stabile. Comportamenti instabili possono comunque essere facilmente riconosciuti quando si applica un certo metodo di approssimazione; nel nostro esempio, basti pensare che il cambiamento di segno tra un passo ed il successivo si verifica $per\ qualsiasi\ valore\ di\ h$, e questa non puo' essere una caratteristica di una funzione a un solo valore.

La (5.19) mostra anche che il metodo di $Eulero$ sara' comunque instabile quando $|1-\alpha h|>1$ (la soluzione approssimata cresce invece di decrescere come la soluzione esatta) ossia, per α positivo, quando $h>\frac{2}{\alpha}$; ricordando che la soluzione esatta va come $exp(-\alpha\cdot x)$, questo porta alla condizione abbastanza ovvia che il passo h deve essere molto minore della scala su cui la soluzione varia significativamente. Questo vale per qualunque metodo approssimato.

Supponiamo ora di avere una equazione del secondo ordine che ha due soluzioni la cui scala di variazione e' molto differente; per fissare le idee, supponiamo che queste soluzioni siano due esponenziali:

$$exp(-\alpha_1\cdot x) \quad , \quad exp(-\alpha_2\cdot x) \quad , \quad \alpha_2 \gg \alpha_1 > 0. \tag{5.25}$$

Supponiamo di voler fissare le condizioni iniziali per $x=0$ e di voler integrare l'equazione per x crescenti. Se ora stiamo cercando la soluzione che varia meno rapidamente, su una scala $\sim \frac{1}{\alpha_1}$, dovremo partire da condizioni iniziali che corrispondano a questa soluzione; anche avendo a disposizione i valori iniziali esatti, quando li forniremo ad un calcolatore dovremo troncarli ad un numero finito di cifre. Questi valori troncati corrisponderanno ad una soluzione che e' una combinazione lineare delle due soluzioni in (5.25); per quanto il coefficiente di $exp(-\alpha_2\cdot x)$ sia piccolo, se il nostro metodo e' instabile per questa soluzione, dopo un certo numero di passi l'instabilita' mascherera' completamente la vera soluzione dell'equazione. Di conseguenza bisognera' scegliere un passo h che sia molto minore di $\frac{2}{\alpha_2}$, ossia della scala di variazione della funzione che cambia piu' rapidamente.

Per esercizio si consiglia di riflettere ai problemi che possono presentarsi nel risolvere le equazioni tipiche della meccanica quantistica, in cui si hanno due soluzioni, una crescente e l'altra decrescente, e solo una ha significato fisico.

5.6 Errore globale del metodo di Eulero

Calcoleremo ora l'errore globale del metodo di $Eulero$. Le ipotesi che vanno fatte sulla funzione $f(x,y)$ saranno evidenti nel corso della dimostrazione e non le espliciteremo. Indichiamo con e_n l'errore in x_n: $e_n = y(x_n) - y_n$.

sottraendo (5.8) da (5.6):

$$\begin{aligned} e_{n+1} &= e_n + h\cdot[y'(x_n) - f(x_n,y_n)] + \frac{1}{2}h^2\cdot y''(\xi) \\ &= e_n + h\cdot[f(x_n,y(x_n)) - f(x_n,y_n)] + \frac{1}{2}h^2\cdot y''(\xi) \\ &= e_n + h\cdot f_y(x_n,\overline{y}_n)\cdot(y(x_n)-y_n) + \frac{1}{2}h^2\cdot y''(\xi) \\ &= e_n + h\cdot f_y(x_n,\overline{y}_n)\cdot e_n + \frac{1}{2}h^2\cdot y''(\xi) \quad , \quad \overline{y}_n \in [y_n, y(x_n)] \end{aligned} \tag{5.26}$$

Indichiamo con L e M i valori massimi assunti da $|f_y|$ e $|y''(x)|$; avremo:

$$\begin{aligned} |e_{n+1}| &\leq |e_n| + h\cdot L\cdot|e_n| + \frac{1}{2}\cdot h^2\cdot M \\ |e_{n+1}| &\leq (1+h\cdot L)|e_n| + \frac{1}{2}\cdot h^2\cdot M \end{aligned} \tag{5.27}$$

Consideriamo la corrispondente equazione alle differenze nella incognita ξ_n:

$$\xi_{n+1} = (1+h\cdot L)\cdot\xi_n + \frac{1}{2}\cdot h^2\cdot M \tag{5.28}$$

che ha per soluzione generale:

$$\xi_n = c \cdot (1 + h \cdot L)^n - \frac{1}{2} \cdot h \cdot \frac{M}{L} \tag{5.29}$$

assegniamo la condizione iniziale $\xi_0 = 0$, che fissa il valore di c:

$$\xi_n = \frac{1}{2} \cdot h \cdot \frac{M}{L} \cdot [(1 + h \cdot L)^n - 1] \tag{5.30}$$

Si puo' ora mostrare, per induzione, che:

$$\xi_n \geq |e_n| \qquad\qquad \forall n \tag{5.31}$$

infatti $\xi_0 \geq |e_0| = 0$ (supponiamo che la condizione iniziale sia $esatta$; stiamo dunque trascurando i problemi di troncamento discussi prima) e:

$$\begin{gathered} \xi_n \geq |e_n| \Rightarrow \\ \Rightarrow (1 + h \cdot L) \cdot \xi_n + \tfrac{1}{2} \cdot h^2 \cdot M \geq (1 + h \cdot L) \cdot |e_n| + \tfrac{1}{2} \cdot h^2 \cdot M \geq |e_{n+1}| \Rightarrow \\ \Rightarrow \xi_{n+1} \geq |e_{n+1}| \end{gathered}$$

(utilizzando la (5.27)). quindi:

$$\begin{aligned} |e_n| &\leq \frac{1}{2} \cdot h \cdot \frac{M}{L} \cdot [(1 + h \cdot L)^n - 1] \leq \\ &\leq \frac{1}{2} \cdot h \cdot \frac{M}{L} \cdot [e^{nhL} - 1] = \\ &= \frac{1}{2} \cdot h \cdot \frac{M}{L} \cdot [e^{(x_n - x_0)L} - 1] \end{aligned} \tag{5.32}$$

(abbiamo usato la disuguaglianza : $1 + hL \leq e^{hL}$).

Come si vede, l'errore globale e' un $O(h)$; questo e' vero in generale per tutti i metodi di approssimazione: l'errore globale e' di un grado inferiore rispetto a quello locale.

5.7 Appendice - Equazioni alle differenze finite

Un'equazione alle differenze finite si scrive, nella sua forma piu' generale:

$$\Phi(y_n, y_{n+1}, y_{n+2}, y_{n+3}, ..., y_{n+k}) = 0 \qquad\qquad \forall \ \ n \tag{5.33}$$

risolvendo rispetto a y_{n+k}:

$$y_{n+k} = f(y_n, y_{n+1}, y_{n+2}, y_{n+3}, ..., y_{n+k-1}) \tag{5.34}$$

k, la differenza tra l'indice piu' alto e quello piu' basso e' detto $ordine$ dell'equazione. L'equazione e' detta lineare quando Φ (o f) e' lineare in ciascuno degli y_i. Le (5.33) e (5.34) costituiscono delle relazioni di ricorrenza tra gli elementi di una successione , y_i; una volta fissati, arbitrariamente, i valori di $k - 1$ elementi (i primi, nel caso piu semplice) di questa successione, esse permettono di calcolare iterativamente tutti gli altri.

Intendiamo per $soluzione$ della equazione una espressione analitica compatta che permette di calcolare qualsiasi y_i senza dover applicare ripetutamente la (5.34). I metodi di soluzione delle

equazioni alle differenze finite lineari a coefficienti costanti (cioe' indipendenti da n) procedono in stretta analogia con quelli usati per le equazioni differenziali lineari a coefficienti costanti; in luogo delle soluzioni esponenziali di queste ultime vanno considerate soluzioni del tipo:

$$y_n \sim \lambda^n \tag{5.35}$$

dove λ e' da determinare. Illustreremo il metodo applicandolo ad una equazioni considerata nel testo; l'equazione:

$$y_{n+1} = (1 + h \cdot L) \cdot y_n + \frac{1}{2} \cdot h^2 \cdot M \tag{5.36}$$

e' del primo ordine, con un termine non omogeneo costante. Per risolverla si considera dapprima l'equazione omogenea associata:

$$y_{n+1} = (1 + h \cdot L) \cdot y_n \tag{5.37}$$

sostituendo $y_n = a \cdot \lambda^n$ si ottiene l'$equazione\ caratteristica$ (algebrica) che permette di determinare i valori di λ non nulli:

$$a \cdot \lambda^{n+1} = a \cdot (1 + h \cdot L) \cdot \lambda^n \Rightarrow \lambda = 1 + h \cdot L \quad , \quad a\ \ arbitrario \tag{5.38}$$

la soluzione $generale$ della equazione omogenea e' dunque $y_n = a \cdot (1 + h \cdot L)^n$. A questa va aggiunta una soluzione particolare della equazione non omogenea; nel caso di termine inomogeneo costante, una possibile soluzione e' una costante, che indichiamo con B:

$$B = (1 + h \cdot L) \cdot B + \frac{1}{2} \cdot h^2 \cdot M \Rightarrow B = -\frac{1}{2} \cdot h \cdot \frac{M}{L} \tag{5.39}$$

la soluzione generale della equazione non omogenea e' dunque data da:

$$y_n = c \cdot (1 + h \cdot L)^n - \frac{1}{2} \cdot h \cdot \frac{M}{L} \tag{5.40}$$

c e' una costante arbitraria che puo' essere fissata imponendo le condizioni iniziali (ad esempio il valore di y_0).

Nel paragrafo (5.5) si puo' trovare un esempio di soluzione di una equazione del secondo ordine.

Nel caso di radici complesse dell'equazione caratteristica, si possono scrivere, come per le equazioni differenziali, combinazioni lineari delle soluzioni complesse che permettono di scrivere la soluzione generale in forma reale; si lascia come esercizio la dimostrazione che la soluzione generale della equazione:

$$y_{n+2} - 2 \cdot y_{n+1} + 2 \cdot y_n = 0 \tag{5.41}$$

e' data da:

$$y_n = (\sqrt{2})^n \cdot \left(c_1 \cdot cos\frac{n\pi}{4} + c_2 \cdot sin\frac{n\pi}{4} \right) \tag{5.42}$$

Capitolo 6

MINIMIZZAZIONE DI FUNZIONI

6.1 Minimizzazione in una dimensione

Il problema puo' essere risolto cercando lo zero della derivata prima della funzione (e stabilendo poi di che tipo di estremo si tratta). Possono essere dunque ripetute tutte le considerazioni fatte nel capitolo sulla ricerca dello zero di una funzione; ad esempio, il metodo di $Newton$ da' la formula di ricorrenza:

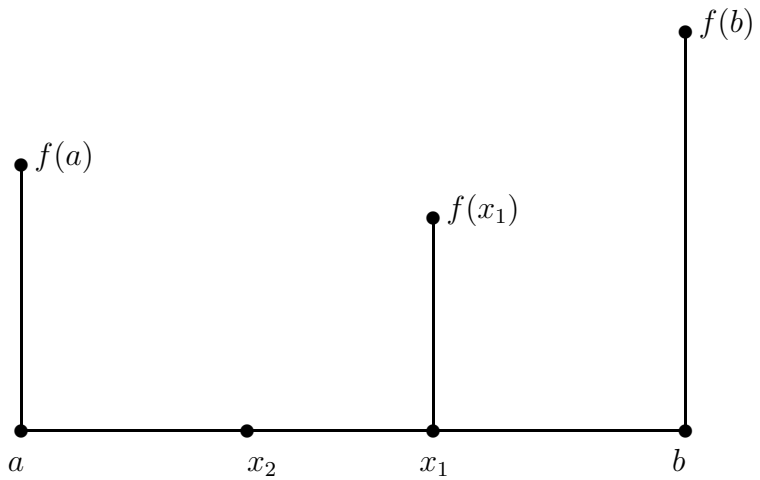
$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)} \tag{6.1}$$

Se tuttavia la forma analitica delle derivate prima e seconda non e' nota esse vanno stimate numericamente; questo comporta il calcolo della funzione piu' volte ad ogni passo.

I metodi che seguono si applicano invece direttamente alla funzione e richiedono un solo calcolo ad ogni passo; discuteremo il metodo della $sezione\ aurea$, che e' il corrispondente del metodo di $bisezione$, ed un metodo $locale$: il metodo di $interpolazione$.

6.1.1 Metodo della sezione aurea

Il Metodo della $sezione\ aurea$ e' un metodo $non\text{-}locale$; si differenzia dal metodo di bisezione nel fatto che per la ricerca di un minimo e' necessario, ad ogni passo, conoscere il valore della funzione in tre punti e calcolarlo in un quarto.



Supponiamo che la funzione $f(x)$ sia continua e di partire da un'intervallo $[a, b]$ in cui esiste

almeno un minimo di $f(x)$. L'esistenza di questo minimo e' assicurata dal fatto che ci sara' qualche altro punto $x_1 \in [a,b]$ tale che $f(x_1) < f(a)$ e $f(x_1) < f(b)$; quindi a, x_1, b saranno i nostri punti di partenza. Si calcola la funzione in un altro punto $x_2 \in [a,b]$: se $f(x_2) < f(x_1)$ il successivo intervallo di ricerca sara' $[a, x_1]$ e a, x_2, x_1 saranno i tre punti di partenza; se $f(x_2) > f(x_1)$ l'intervallo sara' $[x_2, b]$ e x_2, x_1, b i tre punti di partenza; se $f(x_2) = f(x_1)$ bisognera' calcolare $f(x)$ in qualche altro punto per decidere in quale dei due intervalli continuare la ricerca.

Per stabilire le posizioni di x_1 e x_2 richiediamo che il fattore di riduzione, t, sia uguale nei due casi e costante ad ogni passo; si deve dunque avere:

$$x_1 - a = b - x_2 = t \cdot (b - a) \quad , \quad x_2 - a = t \cdot (x_1 - a) \tag{6.2}$$

da cui si ha:

$$t^2 = 1 - t \tag{6.3}$$

che ha come unica soluzione positiva:

$$t = \frac{\sqrt{(5)} - 1}{2} \simeq 0.618 \tag{6.4}$$

se l'intervallo successivo e' $[x_2, b]$ si ottiene lo stesso risultato: $b - x_1 = t \cdot (b - x_2)$, con t dato dalla (6.4). t definisce la $sezione\ aurea$ di un segmento; questa e' stata molto usata, in passato, in architettura, per stabilire le proporzioni tra diverse parti di un edificio.

6.1.2 Metodo di interpolazione

E' un metodo $locale$ equivalente, sotto certi aspetti, al metodo della $secante$. Sono necessari tre punti di partenza ; si cerca il minimo della parabola che interpola la funzione in questi tre punti e si sostituisce questo punto a quello dei tre precedenti in cui la funzione assume il valore piu' grande; si ripete poi la ricerca. Ovviamente questo metodo non funziona se la parabola ha un massimo oppure se i tre punti da cui la parabola deve passare sono quasi allineati (problema analogo a quelli di tutti i metodi $locali$).

6.2 Minimizzazione in piu' dimensioni

Anche se in piu' dimensioni il problema non e' formalmente dissimile da quello in una dimensione, la sua soluzione diventa molto piu' onerosa dal punto di vista del tempo di calcolo ; la formula di $Newton$ (6.1) diventa in questo caso (esercizio!):

$$\underline{x}_{n+1} = \underline{x}_n - \underline{G}_n^{-1} \cdot \underline{g}_n \tag{6.5}$$

dove $\underline{x}$ e' una matrice colonna che rappresenta un punto dello spazio in cui e' definita la funzione, $\underline{g}_n$ e' la matrice colonna delle derivate della funzione calcolate in $\underline{x}_n$ e $\underline{G}_n$ e' la matrice delle derivate seconde della funzione calcolate in $\underline{x}_n$.

I problemi dovuti alla $localita'$ del metodo si manifestano in questo caso nella possibilita' che $\underline{G}$ possa non essere definita positiva in $\underline{x}_n$; in tal caso il passo di $Newton$ porta ad allontanarsi dal minimo.

Ma un problema ulteriore e' dovuto al fatto di dover calcolare ed invertire $\underline{G}$: in molte dimensioni questo puo' essere fatto solo numericamente e richiede un tempo di calcolo notevole rispetto alle semplici operazioni richieste dal metodo unidimensionale.

Una possibile soluzione a quest'ultimo problema puo' essere costituita dall'eseguire un ciclo di N [1] minimizzazioni unidimensionali lungo direzioni opportunamente scelte e nel ripetere piu' volte tale ciclo fino a che il punto finale non soddisfa opportune condizioni.

La scelta piu' banale che si puo' fare e' quella della *minimizzazione lungo gli assi coordinati*, ma ovviamente, non tenendo in alcun conto le caratteristiche della funzione, essa puo' essere molto inefficiente.

Altri metodi modificano gli assi lungo cui eseguire la minimizzazione dopo ogni ciclo: ad esempio scegliendo un asse lungo la direzione che unisce i punti iniziale e finale di un ciclo e gli altri nel piano perpendicolare a tale direzione.

Un'altra possibilita' e' quella di scegliere come direzione lungo cui minimizzare quella del *gradiente* della funzione in un punto. Trovato il minimo lungo questa direzione, si minimizza lungo il gradiente in questo nuovo punto, e cosi' via.

Per ciascuno di questi metodi e' facile immaginare qualche funzione per la quale esso sara' inefficiente (vedi le figure alla fine delle note). In particolare si potrebbe pensare che la scelta di minimizzare lungo i *gradienti* sia particolarmente felice; ma il gradiente in un punto da' la direzione di massima variazione *locale* della funzione, e tale direzione non e' detto che abbia qualcosa a che vedere con la direzione in cui si trova il minimo della funzione.

Il metodo dei *gradienti coniugati*, invece, permette di individuare N direzioni per le quali un ciclo di N minimizzazioni risulta equivalente ad un passo di $Newton$ (6.5). In tal modo si ottiene la stessa efficienza del metodo di $Newton$ senza dover nemmeno *calcolare* la matrice delle derivate seconde.

Nei paragrafi successivi, prima di descrivere quest'ultimo metodo, discuteremo il metodo del *simplesso*, che e' l'equivalente in piu' dimensioni dei metodi *non-locali* visti in precedenza.

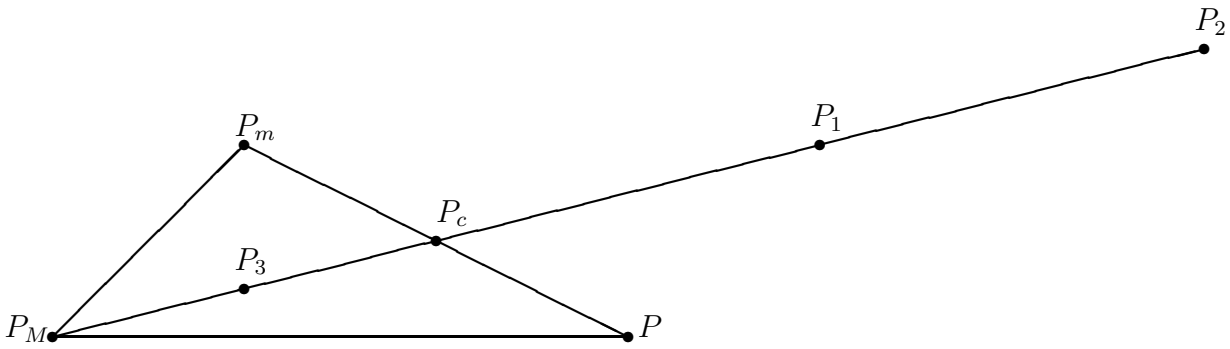
6.2.1 Metodo del simplesso

In uno spazio ad N dimensioni un *simplesso* e' definito come la figura geometrica a $N+1$ vertici: triangolo e tetraedro in due e tre dimensioni.

Il metodo consiste nello scegliere $N+1$ punti in modo che il simplesso da essi formato racchiuda o sia di dimensioni comparabili a quelle della regione in cui si cerca il minimo.

Nella figura seguente il simplesso di partenza, in uno spazio a due dimensioni, e' definito dai punti P_M, P_m, P. P_M e P_m sono i punti in cui la funzione assume il valore massimo e minimo rispettivamente. Il metodo consiste nell'applicazione di un algoritmo che permetta ad ogni passo di sostituire il punto P_M con un nuovo punto in cui la funzione assume valore piu' basso, ed anche, in certe condizioni, di restringere o dilatare le dimensioni del simplesso iniziale. Questo algoritmo permettera' di spostare il simplesso verso la regione del minimo e, quando questo minimo sara' vicino, di restringerlo attorno ad esso.

[1] indichiamo con N la dimensione dello spazio in cui la funzione e' definita.



Un algoritmo che abbia queste caratteristiche non e' unico, e puo' essere piu' o meno efficiente; ne descriviamo uno piuttosto semplice:

- si individua P_M e si trova il baricentro, P_c, degli altri $N-1$ punti; si calcola la funzione in P_1 di coordinate [2] :

$$\underline{x}_1 = \underline{x}_M + 2 \cdot (\underline{x}_c - \underline{x}_M) \tag{6.6}$$

  Se $f(\underline{x}_m) < f(\underline{x}_1) < f(\underline{x}_M)$ si sostituisce P_M con P_1.

  Se $f(\underline{x}_1) < f(\underline{x}_m)$ si calcola ancora la funzione in P_2 di coordinate:

$$\underline{x}_2 = \underline{x}_M + 3 \cdot (\underline{x}_c - \underline{x}_M) \tag{6.7}$$

  e si sostituisce P_M col migliore tra P_1 e P_2.

  Se invece $f(\underline{x}_1) > f(\underline{x}_M)$ si calcola la funzione in un punto P_3 piu' vicino a P_M:

$$\underline{x}_3 = \underline{x}_M + \frac{1}{2} \cdot (\underline{x}_c - \underline{x}_M) \tag{6.8}$$

  se nemmeno in questo punto la funzione assume un valore inferiore a $f(\underline{x}_M)$, si contraggono tutte le dimensioni del simplesso di uno stesso fattore, lasciando fisso il punto P_m.

Dopo aver ridefinito il simplesso, si ripetere la procedura fino a che le sue dimensioni non sono dell'ordine della precisione richiesta.

6.2.2 Direzioni coniugate - Gradienti coniugati

Il metodo di $Newton$ minimizza esattamente, in un solo passo, una forma quadratica; questa puo' essere sempre scritta nel modo seguente:

$$F(\underline{x}) = F(\underline{0}) + \underline{x}^T \cdot \underline{g} + \frac{1}{2} \cdot \underline{x}^T \cdot \underline{G} \cdot \underline{x} \tag{6.9}$$

dove $\underline{0}$ rappresenta l'origine delle coordinate; la matrice dei gradienti, $\underline{g}$ e quella delle derivate seconde, $\underline{G}$, vanno calcolate in $\underline{0}$. Per ipotesi $F(\underline{x})$ ha un minimo, quindi $\underline{G}$ e' definita positiva.

Vedremo ora che e' possibile, eseguendo N minimizzazioni unidimensionali lungo le *direzioni coniugate* che definiremo, ottenere lo stesso risultato, cioe' il minimo *esatto* della forma quadratica.

[2] indicheremo con $\underline{x}$ la matrice delle coordinate dei vari punti

Un sistema di direzioni coniugate rispetto alla matrice $\underline{G}$ e' definito come un insieme di N vettori, $\underline{d}_i$, in generale non ortogonali, linearmente indipendenti che soddisfano la proprieta' [3]:

$$\underline{d}_i^T \cdot \underline{G} \cdot \underline{d}_j = \delta_{ij} \tag{6.10}$$

Dati i vettori $\underline{d}_i$ potremo scrivere ogni vettore $\underline{x}$ come combinazione lineare:

$$\underline{x} = \sum_{i=1}^{N} \lambda_i \cdot \underline{d}_i \tag{6.11}$$

le λ_i rappresentano le coordinate del vettore $\underline{x}$ nel sistema di riferimento che ha come sistema di base i vettori $\underline{d}_i$.

Evidentemente (esercizio!) la funzione $F(\underline{x})$ potra' scriversi come somma di N funzioni di una variabile:

$$F(\underline{x}) = \sum_{i=1}^{N} f_i(\lambda_i) \tag{6.12}$$

quindi N minimizzazioni lungo le direzioni $\underline{d}_i$,ossia rispetto alle variabili λ_i, porteranno al minimo della funzione $F(\underline{x})$.

Il problema e' ora trovare un'insieme di direzioni coniugate; la (6.10) e' soddisfatta dagli autovettori della matrice $\underline{G}$ (e di fatto la (6.12) e' una diagonalizzazione della forma quadratica). La ricerca degli autovettori richiede tuttavia l'inversione di $\underline{G}$, quindi la complessita' del calcolo e' maggiore che nell'applicazione diretta del netodo di $Newton$.

Le direzioni coniugate non sono pero' uniche: esiste un metodo, simile a quello di *ortogonalizzazione di Schmidt*, che permette di ottenere un insieme di direzioni coniugate a partire da un vettore arbitrario; questo metodo, di cui non scriviamo le formule, richiede solo il calcolo di $\underline{G}$ ma non la sua inversione.

Un terzo metodo per la ricerca di direzioni coniugate e' quello dei *gradienti coniugati* che richiede solo una serie di minimizzazioni unidimensionali lungo direzioni calcolate ogni volta a partire dai gradienti della funzione nei punti di minimo. Ossia si parte da un punto arbitrario, si minimizza lungo la direzione del gradiente in quel punto; la direzione successiva si calcolera' a partire dal gradiente nel punto di minimo, e cosi' via. La formula che definisce le N direzioni coniugate e' la seguente, e la diamo senza dimostrazione:

$$\underline{d}_{i+1} = -\underline{g}_{i+1} + \frac{\underline{g}_{i+1}^T \cdot \underline{g}_{i+1}}{\underline{g}_i^T \cdot \underline{g}_i} \cdot \underline{d}_i \tag{6.13}$$

[3] usiamo gli stessi simboli sottolineati per indicare un vettore o la matrice colonna delle sue coordinate nel sistema di riferimento utilizzato