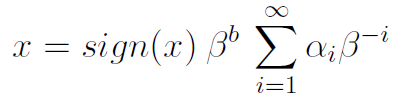
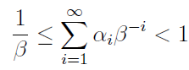


**Rappresentazione dei numeri**

Scelto un qualunque numero intero β > 1, ogni numero non nullo x ε R ammette una rappresentazione in base β:

dove :

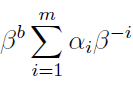
* b ε Z viene chiamato esponente**.**
* αi [0, β-1] vengono chiamate cifre.
* sign(x) è una funzione che assume:
  + valore 1 se x > 0.
  + valore -1 se x < 0.
* La sommatoria di tutti gli αi β-i è detta mantissa del numero x. Si verifica che:

**Teorema di rappresentazione**

Data una base intera β > 1 e un qualunque numero reale x diverso da 0, esiste un’unica rappresentazione in base β tale che:

1. sia α1 ≠ 0
2. non vi sia un intero k tale che: αj = β – 1 ∀ p j > k

Questo vuol dire che fissato β sono univocamente determinati i numeri b e α*i* .

All’interno di un calcolatore si possono rappresentare solo un certo numero **m** di cifre della mantissa di x. Quindi la rappresentazione di x su un calcolatore può essere:

* y = tr(x) =



* y = rd(x) =

dove:

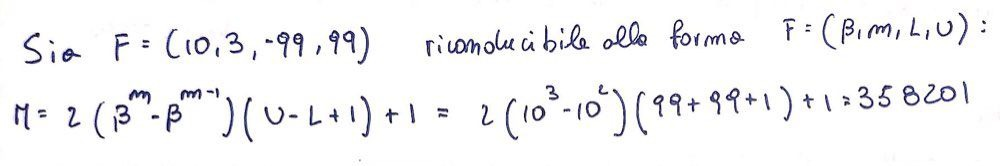
* tr(x) si chiama rappresentazione per troncamento del numero reale x.
* rd(x) si chiama approssimazione per arrotondamento che, in generale, è una migliore approssimazione del numero reale x.

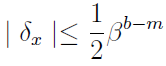
**Cardinalità dei numeri di macchina**

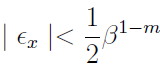
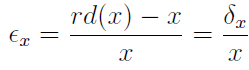
Si indichi con M l’insieme (finito) dei numeri z rappresentabili all’interno di un calcolatore, comunemente chiamati numeri di macchina, fissati β, m, L, U la cardinalità di M risulta essere:



Spessa l’insieme M viene indicato con il simbolo F (β, m, L, U ).

Esempio:

**Errore assoluto**

**Errore relativo**

Il numero u = ½ β1-m si dice **precisione di macchina** ed è il massimo errore relativo commesso nel passaggio da x a rd(x). Quando l’errore di approssimazione non super ½ b1-k si dice che l’approssimazione è corretta almeno fino alla k-esima cifra significativa.

**Operazioni di macchina**

Nell’insieme M non tutte le proprietà delle quattro operazioni elementari risultano verificate, in quanto il risultato di un’operazione deve essere ricondotto ad un numero di macchina ( ad esempio l’addizione non gode della proprietà associativa).

**Errore nel calcolo di una funzione**

Sia f : Rn -> R una funziona ad n variabili e a valori reali. Si vuole calcolare f(P) = f(x1, x2,…,xn) in un punto assegnato P0 = (x(0)1,…,x(0)n). In generale non si ottiene il valore f(P0) cercato a causa delle approssimazioni che si introduco. Queste sono di due generi:

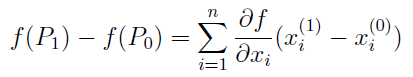
1. le operazioni aritmetiche che compaiono in f(P) devono essere sostituite con le corrispondenti operazioni macchina. Ciò equivale a sostituire la funzione f(P) con un’altra funzione fa(P).
2. le coordinate del punto P0 possono non essere esattamente rappresentabili e quindi si devono approssimare ( basti pensare al punto P0 = ( √2, π) ).

Quindi possiamo concludere dicendo che, volendo calcolare la funzione f(P0) siamo costretti a calcolare il valore fa(P1). In altre parole approssimiamo il valore della funzione nel punto assegnato cambiando la funzione e il punto.

**Errore assoluto**

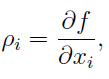
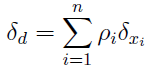
Assegnato il punto P0 ε Rn e la funzione f. Vogliamo calcolare f(P). Consideriamo l’insieme:

dove ai, bi ε R e si supponga che ai ≤ x(0)i ≤ bi. D si dice **insieme di indeterminazione** del punto P0 . l valore f(P0) viene sostituito dal valore calcolato fa(P1) con P1 ε D e quindi l’errore commesso risulta essere fa(P1) – f(P0). Di questo errore (detto errore totale) possiamo dare una stima. Poniamo:

* fa(P1) – f(P1) è detto **errore assoluto algoritmico.** Dipende dall’aver sostituito la funzione f(P) con la funzione fa(P). Questo è definito e stimabile.
* f(P1) – f(P0) è detto **errore assoluto trasmesso dai dati**. Questo non è stimabile ma si può dare una rappresentazione generale tramite la formula di Taylor:

Indichiamo quindi:

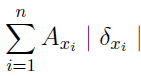
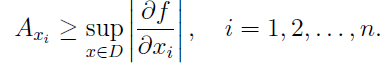
* δf = fa(P1) – f(P0) errore totale
* δa = fa(P1) – f(P1) errore algoritmico
* δd = f(P1) – f(P0) errore trasmesso dai dati

Risulta quindi che: δf = δa + δd

Ponendo poi: δxi = x(1)i – x(0)i , l’equazione di Taylor per l’errore trasmesso dai dati diventa:

I valori ρi coefficienti di amplificazione degli errori δxi . Una limitazione per il modulo dell’errore assoluto è:

|fa(P1) – f(P0)| ≤ Ea + Ed

* Ea è il massimo modulo dell’errore assoluto dovuto al particolare algoritmo usato.
* Ed = con

**Problema diretto**

Se si conosce una stima dell’errore algoritmico e degli errori δxi nonché le Axi , si può stabilire un confine superiore per l’errore assoluto con cui si è calcolata la funzione nel punto.

Immagine che contiene testo, finestra

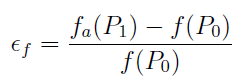
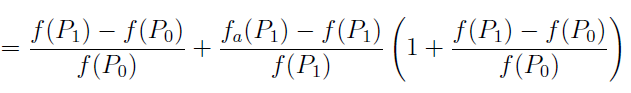
Descrizione generata automaticamente

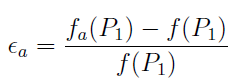
**Problema inverso**

Immagine che contiene tavolo

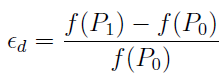
Descrizione generata automaticamenteConsiste nel richiedere che il valore fa(P1) sia tale che l’errore assoluto |fa(P1) – f(P0)|risulti ≤ ad un valore prefissato, per cui si deve cercare sia un algoritmo opportuno fa(P) sia un punto P1 che soddisfano la richiesta.

**Errore relativo**

L’errore relativo che si commette nel calcolo di una funzione f(P) in un punto assegna P0 è:

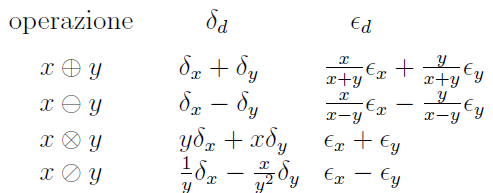
Come abbiamo fatto prima, dividiamo:

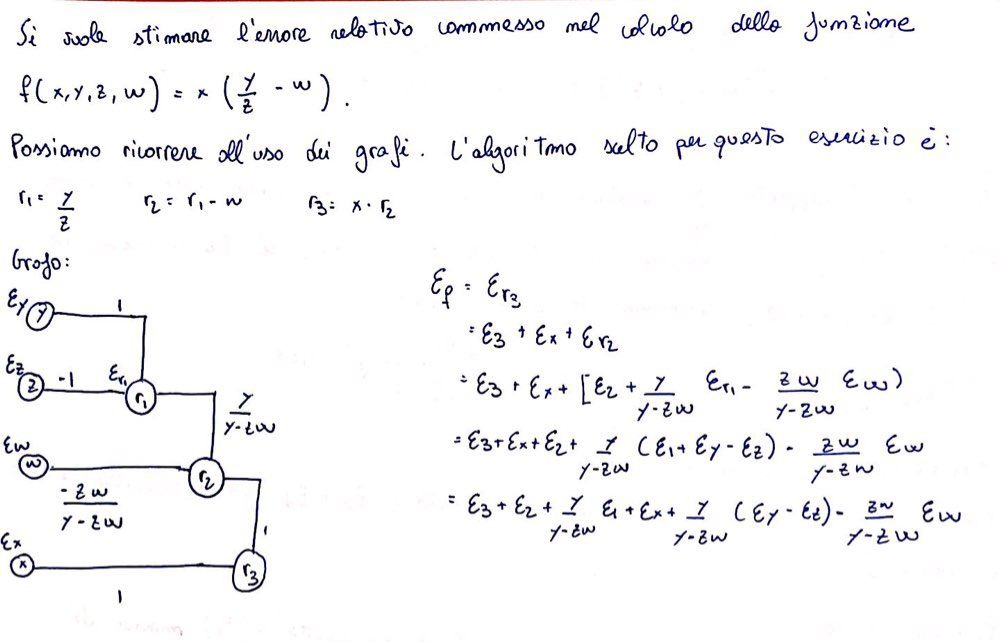
* Errore relativo algoritmico:



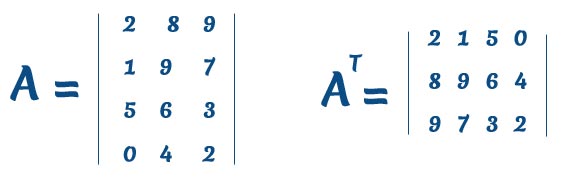
* Errore relativo trasmesso dai dati:

Si ottiene l’errore totale: εf = εa + εd + εaεd

**Gli errori nelle quattro operazioni**

**Complementi ed esempi:**

**Matrice trasposta**

Data la matrice A ε Cm\*n, la matrice B ε Cn\*m i cui elementi sono bi,j = aj,i si dice matrice trasposta.

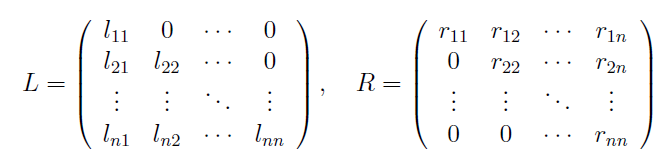
**Matrice trasposta coniugata**

Data la matrice A ε Cm\*n, la matrice B ε Cn\*m i cui elementi sono bi,j = j,i si dice matrice trasposta coniugata.

**Matrice diagonale**

È una matrice D quadrata con gli elementi non diagonali nulli. Si scrive anche D = diag(d1, d2,…,dn).

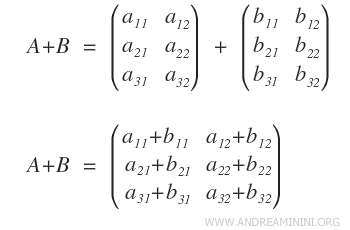
**Matrice triangolare**

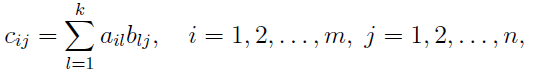
Sono matrici quadrate del tipo:

L = triangolare inferiore

R = triangolare superiore

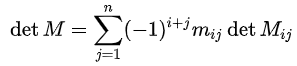
**Operazioni tra matrici**

**Somma**: C = A + B => ci,j = ai,j + bi,j . Valgono le proprietà associativa e commutativa.

**Moltiplicazione:** Data A ε Cm\*k , B ε Ck\*n , la matrice prodotto C = A \* B si definisce ponendo:

e risulta quindi C ε Cm\*n . La moltiplicazione gode della proprietà associativa e distributiva ma non commutativa infatti è sempre bene distinguere pre-moltiplicazione da post-moltiplicazione.

**Determinante**

Per il calcolo del determinante si può fare uso del primo teorema di Laplace:

Una matrice A si dice **singolare** se det(A) = 0, **non singolare** se det(A) ≠ 0.

**Minore di ordine k**

Data una matrice A ε Cm\*n ed un intero k ≤ min {m,n}, si dice minore di ordine k il determinante di una matrice ottenuta da A prendendo tutti gli elementi sulla intersezione di k righe e k colonne comunque fissate.

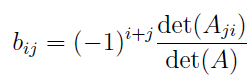
**Minore principale di ordine k**

Data una matrice A ε Cn\*n si dicono minori principali di ordine k i determinanti delle sottomatrici di ordine k estratte da A e aventi diagonale principale composta da elementi della diagonale principale di A.

**Minore principale di testa**

Data una matrice A ε Cn\*n si dice minore principale di testa di ordine k il determinante della sottomatrice di ordine k formata dalle prime k righe e k colonne di A.

**Matrice inversa**

Data una matrice a ε Cn\*n non singolare, si dice matrice inversa di A la matrice B ε Cn\*n :

Il determinante di una matrice A è il reciproco del determinante della sua inversa (det(A-1) = 1/det(A) ).

**Matrici particolari**

**Matrice hermitiana:** matrice che coincide con la sua trasposta coniugata. A = AH. Nel campo dei reali coincide con la definizione di matrice simmetrica.

**Matrice antihermitiana:** matrice che coincide con la sua trasposta coniugata cambiata di segno. A = -AH

**Matrice normale:** matrice che commuta la sua matrice trasposta coniugata. AHA = AAH

**Matrice simmetrica:** matrice uguale alla sua trasposta. A = AT

**Matrice antisimmetrica:** matrice uguale alla sua trasposta cambiata di segno. A = -AT

**Matrice unitaria:** una matrice è unitaria se è invertibile e la sua inversa è uguale alla sua trasposta coniugata. AHA = AAH = I

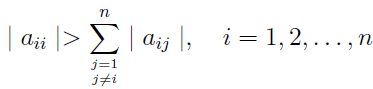
Nel campo reale, una matrice hermitiana è anche simmetrica e le matrici unitarie hanno come inversa la loro trasposta coniugata. Una matrice reale e unitaria si dice ortogonale.

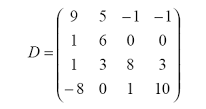
**Matrice di permutazione**

È una matrice ortogonale ottenuta dalla matrice identica operando su di essa una permutazione delle colonne o delle righe. È una matrice ortogonale poiché risulta PTP = PPT = I. Il prodotto di una matrice A per una matrice di permutazione P produce su A una permutazione delle colonne o delle righe, precisamente:

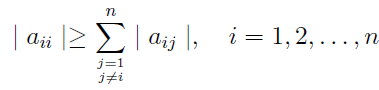
* AP presenta la stessa permutazione di colonne che si è operata su I per ottenere P
* AP presenta la stessa permutazione di righe che si è operata su I per ottenere P

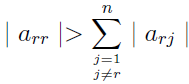
**Predominanza diagonale forte**

Una matrice si dice a predominanza diagonale forte se l’elemento diagonale è maggiore della somma dei moduli degli elementi non diagonali:

Esempio:

**Predominanza diagonale debole**

Una matrice si dice a predominanza diagonale forte se l’elemento diagonale è maggiore o uguale della somma dei moduli degli elementi non diagonali:

e almeno per un indice r, 1 ≤ r ≤ n, si ha:

**Matrice convergente**

Una matrice si dice convergente se: O è la matrice nulla.

**Teorema di Rouchè – Capelli**

Il sistema lineare ammette soluzione se e solo se rango(A) = r( A | b ).

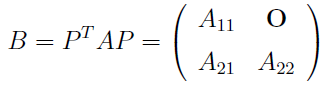
Con (A| b) si intende la matrice completa del sistema costituita da n righe ed n+1 colonne. Se r(A) = n la soluzione è unica mentre se r(A) < n l’insieme delle soluzioni è un sottospazio di Cn di dimensione n-r(A). Se il det(A) ≠ 0 il sistema si dice **normale** e la soluzione si ricava mediante la regola di Cramer. Nel caso particolare b = 0, il sistema Ax = b si dice omogeneo. Se il det(A) ≠ l’unica soluzione è x = 0; segue che un sistema omogeneo ha soluzioni non nulle se e solo se il det(A) = 0.

**Matrici partizionate**

Sono matrici i cui elementi sono sottomatrici. Una qualunque matrice può essere partizionata a blocchi in molti modi, è importante il caso in cui i blocchi diagonali sono quadrati.

**Matrici riducibili**

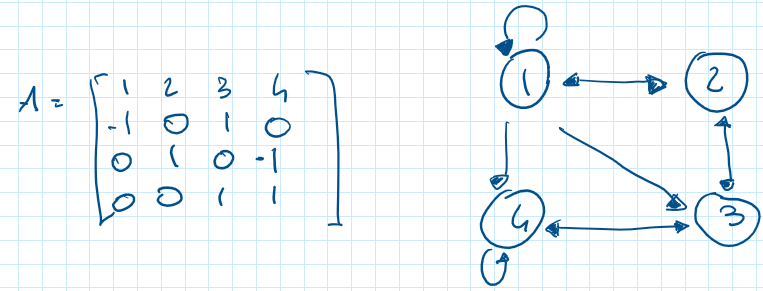
Una matrice si dice riducile se esiste una matrice di permutazione P tale che la matrice PTAP sia della forma:



Con i blocchi diagonali quadrati. La matrice B si ottiene dalla matrice A operando sulle righe e sulle colonne la stessa permutazione operata sulle colonne di I per ottenere P. Per stabilire se esiste o meno una matrice P, servono le definizioni ed il teorema seguente:

**Grafo orientato**

Data una matrice A ε Cn\*n, fissati n punti detti nodi, si dice grafo orientato associato ad A il grafo che si ottiene congiungendo ni a nj con un cammino orientato da ni a nj se ai,j ≠ 0.

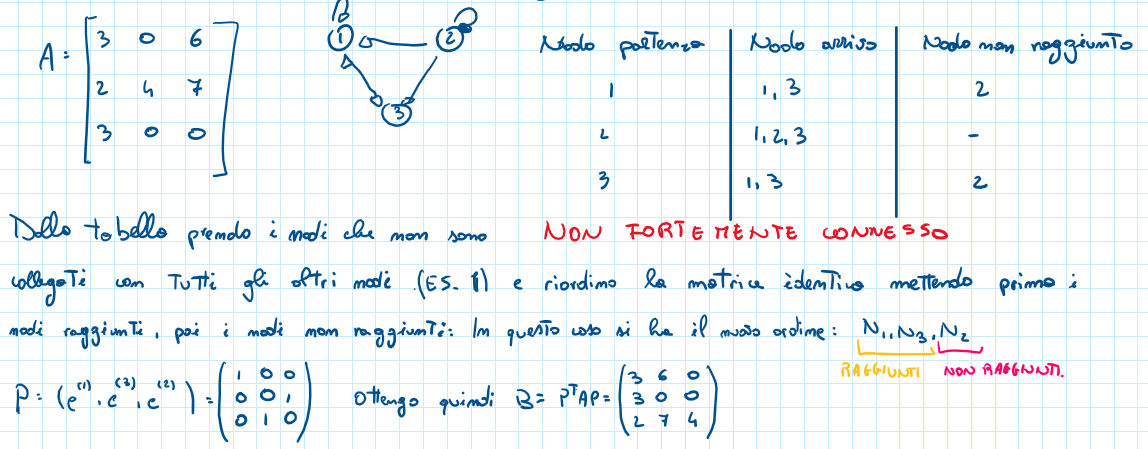
Esempio:

**Grafo fortemente connesso**

Un grafo si dice fortemente connesso se da ogni nodo ni è possibile raggiungere un qualunque altro nodo nj seguendo un cammino orientato eventualmente passando per altri nodi.

Una matrice è irriducibile se il grafo orientato associato ad essa è fortemente connesso.

**Esempio**

****

**Autovalori e autovettori**

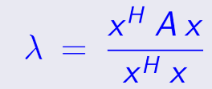
Data una matrice A ε Cn\*n si dice autovalore di A ogni numero λ ε C tale che il sistema lineare Ax = λx, x ε Cn abbia soluzioni x ≠ 0. Risolvere il sistema lineare Ax = λx è la stessa cosa di risolvere il sistema omogeneo (A-λI) x = 0. Ne segue che gli autovalori di A sono tutti e soli i numeri λ che verificano: det(A-λI) = 0. Dal calcolo del determinante si ottiene il polinomio caratteristico:

dove i coefficienti σi sono la somma dei minori principali di ordine i estratti dalla matrice A. In particolare risulta: . σ1 si dice traccia di A e si indica con tr(A).

**Implicazioni importanti:**

* Se esiste un indice ‘i’ tale che λi = 0 allora det(A) = 0.
* Se il det(A) = 0 allora esiste un indice i tale che λi = 0.
* Se λ è autovalore della matrice A, allora λk è autovalore della matrice Ak e gli autovettori di A sono anche gli autovettori di Ak. In particolare, se la matrice A è non singolare, si ha che gli autovalori di A-1 sono i reciproci degli autovalori di A.
* Gli autovalori di una matrice hermitiana sono tutti reali.

**Autovalori di una matrice hermitiana**

Dall’uguaglianza Ax = λx, si ottiene premoltiplicando entrambi i lati per xH : xHAx =xHλx ed ancora dividendo per il numero reale e positivo xHx = Σ|xi| si ottiene:

**Raggio spettrale**

Si dice raggio spettrale della matrice A il numero reale non negativo:

Una matrice è **convergente** se e solo se il suo raggio spettrale è < 1.

**Trasformata per similitudine**

Data una matrice A ε Cn\*n ed una matrice S ε Cn\*n non singolare, si dice trasformata per similitudine della matrice A, la matrice B tale che B = S-1AS . Le matrici A,B si dicono simili (A ~ B). Due matrici simili hanno gli stessi autovalori.

**Matrice diagonalizzabile**

Una matrice A si dice diagonalizzabile se esiste una matrice X non singolare tale che X-1AX = D con D = diag(λ1, λ2,…,λ3). Affinchè una matrice sia diagonalizzabile, deve risultare che α(λ) = γ(λ) (ovvero la molteplicità algebrica deve essere uguale alla molteplicità geometrica). L’uguaglianza è sicuramente verificata se risulta α(λ) = 1 ovvero la matrice A ha autovalori a due a due distinti.

**Traslazione dello spettro**

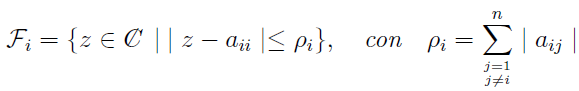
Sia λ autovalore di A e q ε C allora B = A + qI ha come autovalore μ = λ + q con molteplicità algebrica pari a quella di λ e B ha gli stessi autovettori di A.

**Dimostrazione**

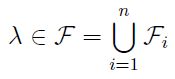
det(B – μl) = det(A + ql – μl) = det (A – I(μ – q) ) da cui λ = μ – q -> μ = λ + q

Le relazioni sugli autovettori si deducono da:

Bx = μx -> (A+Iq) = μx -> Ax = (μ-q)x -> Ax = λx

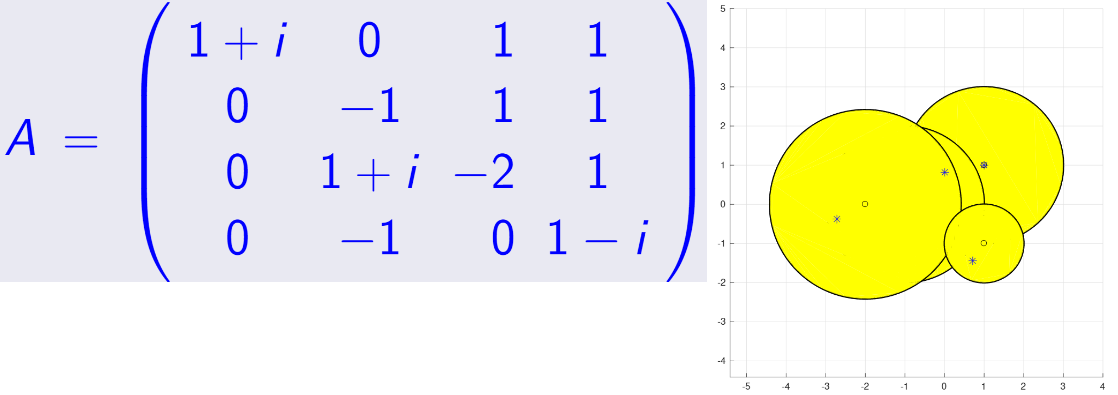
**Primo teorema di Gershgorin:**

sia A ε Cn\*n, si indichino con Fi i = i,2,…,n gli insiemi (chiamati cerchi di Gershgorin) :

Sul piano complesso, il cerchio di Gershgorin Fi ha centro nel punto aii e raggio ρi. Allora se λ è autovalore di A si ha:

L’insieme F a cui appartengono tutti gli autovalori è l’unione dei cerchi Fi detti cerchi di Gershgorin:

**Esempio:**

****

Una conseguenza del primo teorema di Gershgorin è che una matrice a predominanza diagonale forte non è singolare.

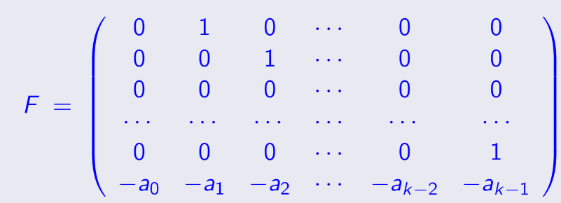
**Secondo teorema di Gershgorin:** se M1 è l’unione di k cerchi di Gershgorin e M2 è l’unione dei n-k rimanenti ed è M1 ∩ M2 = 0, allora k autovalori appartengono a M1 e n-k autovalori appartengono a M2 .

**Corollario:** se una matrice presenta cerchi di Gershgorin a due a due disgiunti allora gli autovalori sono due a due distinti

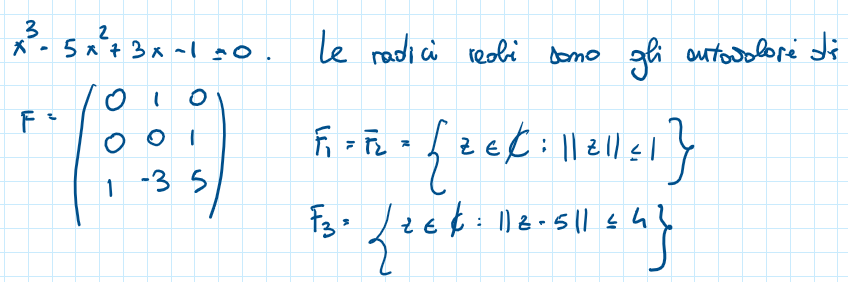
**Terzo teorema di Gershgorin:** sia A irriducibile, se un autovalore appartiene alla frontiera dell’unione dei cerchi di Gershgorin esso appartiene alla frontiera di tutti i cerchi costituenti l’insieme F.

**Corollario:** una matrice a predominanza diagonale debole ed irriducibile è non singolare.

**Matrice di Frobenius**

Sia data l’equazione algebrica: xk + ak-1xk-1 + … + a1x + a0 = 0 ; con ai ε C, si consideri la matrice quadrata di ordine k:

Detta matrice di Frobenius ( o associata o compagna). Si verifica che l’equazione caratteristica è l’equazione della caratteristica della matrice F per i cui i suoi valori sono le radici reali dell’equazione algebrica facendo uso del primo teorema di Gershgorin.

**Esempio:**

Le matrici Hermitiane e semidefinite positive (negative) hanno autovalori non negativi (non positivi) e viceversa.

**Norme**

Sono utili per confrontare due vettori o due matrici.

**Norme vettoriali**

Si indica con ||x|| una funzione definita nello spazio vettoriale Cn a valori non negativi, una funzione che verifica le seguenti condizioni:

1. ||x|| = 0 ⬄ x = 0
2. ||αx|| = |α| ||x||
3. ||x+y || ≤ ||x|| + ||y||

Le norme più usate sono:

**Norma 1)** ||x||1 =

**Norma 2 o norma euclidea)** ||x||2 =

**Norma ∞ )** ||x||∞  =

**Immagine che contiene testo, antenna

Descrizione generata automaticamenteSfera unitaria di R2, per le tre norme classiche**

S = {x ε R2 | ||x|| ≤ 1 }

**Norme matriciali**

Si indica con ||A|| una funzione definita in Cn\*n a valori non negativi, che verifica le seguenti condizioni;

1. ||A|| = 0 ⬄ A = 0
2. ||αA|| = |α| ||A||
3. ||A+B|| ≤ ||A|| + ||B||
4. ||AB|| ≤ ||A|| ||B||

Immagine che contiene testo

Descrizione generata automaticamenteSi possono ottenere norme matriciali facendo uso delle norme vettoriali definendo:

Una norma matriciale si dice compatibile o coerente con una norma vettoriale se si ha ||Ax||≤||a|| ||x||. Le norme indotte sono sempre coerenti. Le norme matriciali indotte dalle 3 norme vettoriali classiche sono:

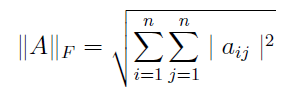
Le norme più usate sono:

**Norma 1)** =

**Norma 2)** =

**Norma ∞) =**

**Norma di Frobenius**

Esistono norme matriciali che non sono indotte da norme vettoriali, ad esempio la norma matriciale di Frobenius:

**Teorema di Hirsh**

Per ogni norma matriciale (indotta o no) vale la relazione ρ(A) ≤ ||A||.

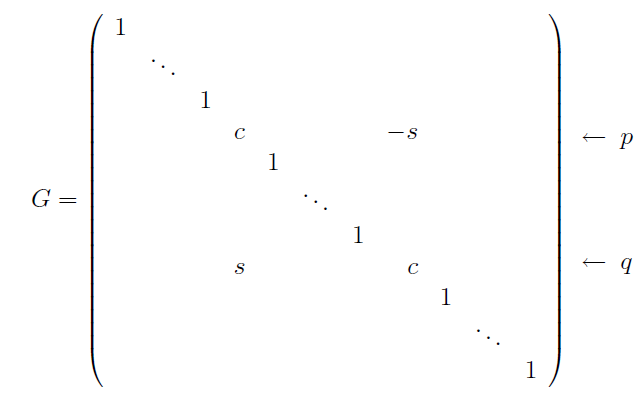
**Dimostrazione**

Sia λ autovalore di A, quindi si ha Ax = λx con x autovettore destro associato a λ. Consideriamo la matrice B = (x|0|0|0…|0) ε Cn\*n per cui vale l’uguaglianza AB = λB. Passando alle norme si ha: ||AB|| = ||λB||. Per le proprietà delle norme matriciali ||A|| ||B|| ≥ |λ|||B|| . La matrice B non è la matrice nulla quindi risulta ||A|| ≥ |λ|.

Da questo deriva direttamente che una matrice per essere convergente, basa che abbia una sua norma minore di 1. Un esempio in cui vale ρ(A) = ||A|| è dato dalle matrici hermitiane (A = AH)

**Matrice di rotazione**

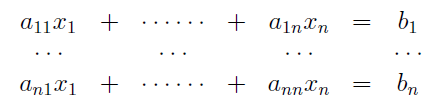
Per le matrici che hanno una forma tipo:



Dove G ε Rn\*n e c2 + s2 = 1 si dicono matrici di rotazione. Nel caso in cui s ≠ 0 gli autovalori sono:

* λ1 = … = λn-2 = 1
* λn-1 = c - √-s2
* λn = c+ √-s2

**Sistemi lineari**

Siano dati una matrice A ε Cn\*n (matrice dei coefficienti) ed un vettore b ε Cn (vettore dei termini noti), un sistema di n equazioni lineari in n incognite si può rappresentare nella forma Ax = b ovvero:

Dove si cerca x ε Cn è il vettore delle incognite che soddisfa tutte le equazioni.

**Sistemi di equazioni lineari**

Esistono due metodi per risolvere sistemi di equazioni lineari, detti:

1. Metodi diretti: si arriva alla soluzione (a meno di errori di arrotondamento) con un numero finito di operazioni sui dati.
2. Metodi iterativi: la soluzione viene approssimata dai termini di una successione di cui la soluzione cercata è il limite.

**Metodi diretti**

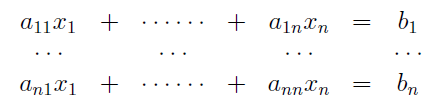
**Cramer**

Immagine che contiene testo

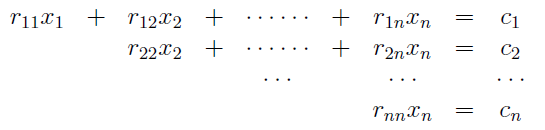
Descrizione generata automaticamenteDato il sistema Ax = b la soluzione si ottiene calcolando:

dove Ai è ottenuta da A sostituendo il vettore b alla i-esima colonna. È sicuramente utile per le dimostrazioni ma non risulta un buon metodo dal punti di vista numerico per colpa del suo elevato costo computazionale.

**Algoritmo base del metodo di Gauss (o metodo di eliminazione)**

Dato il sistema di n equazioni lineare:

Dove i coefficienti ai,j (ai,i ≠ 0 altrimenti l’algoritmo si interrompe) e i termini noti bi sono numeri reali, si cerca un vettore xT = (x1, …, xn ) che verifichi le uguaglianze di sopra. Introdotta la matrice dei coefficienti A (formata da tutti gli ai,j) e il vettore dei termini noti (bi ), il sistema si può scrivere nella forma Ax = b. Si suppone A non singolare per garantire l’esistenza e l’unicità della soluzione. Il metodo di Gauss consiste nel trasformare il sistema Ax = b in un altro sistema Rx = c dove R è una matrice triangolare superiore con ri,i ≠ 0 (in mancanza di questa condizione, l’algoritmo si interrompe). Il nuovo sistema ha una forma del genere:



L’uso di questa tecnica comporta un numero di operazioni, per un sistema di n equazioni in n incognite, pari a n3 / 3 + n2 – n/3 . Le soluzioni sono date da:

Immagine che contiene testo

Descrizione generata automaticamente

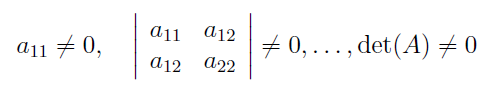
**Immagine che contiene tavolo

Descrizione generata automaticamenteEsempio**

**Immagine che contiene testo

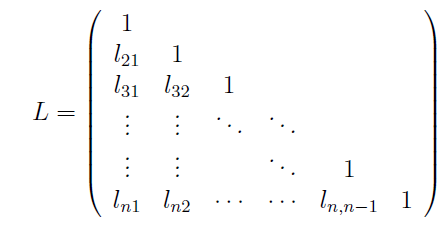
Descrizione generata automaticamenteConfronti dei costi computazionali**

**Fattorizzazione LR**

L’algoritmo può essere considerato come un procedimento che trasforma una data matrice A in una matrice triangolare superiore R. Nell’ipotesi che valgano le seguenti condizioni:

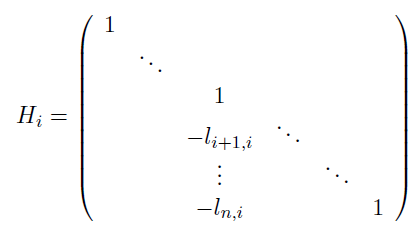
l’algoritmo di eliminazione produce la fattorizzazione A = LR dove:

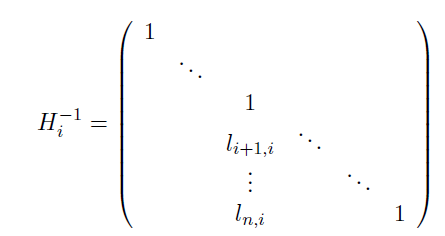
* R è la matrice triangolare superiore data dai coefficienti del sistema.
* L ha la forma:



Gli elementi al di sotto della diagonale principale coincidono con i moltiplicatori dell’algoritmo di eliminazione.

**Dimostrazione**

Siano A1, A2,…,An-1 = R le matrici dei successivi sistemi equivalenti al sistema Ax = b che si ottengono dopo ciascuna eliminazione. Le varie matrici Ai le calcoliamo come: A1 = H1A; A2 = H2A1,…..; An-1 = Hn-1An-2 = R con:

Posto Li = Hi-1 e tenuto conto che

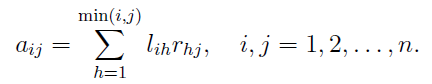
**Immagine che contiene tavolo

Descrizione generata automaticamente**e che L1L2…Ln-1 = L segue che Hn-1Hn-2…H1A = R. Nel caso di una matrice A qualunque.

**Metodo di fattorizzazione**

La conoscenza di una fattorizzazione della matrice A è immediatamente utile per scrivere il sistema Ax = b nella nuova forma LRx = b che ci riconduce immediatamente alla risoluzione dei due sistemi triangolari:

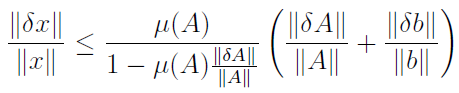
* Lc = b
* Rx = c

Una conseguenza di questa fattorizzazione è: det(A) = det(R).do Se lo scopo della fattorizzazione è la risoluzione dei sistemi si preferisce costruire direttamente le matrici L ed R. Si hanno così i metodi di fattorizzazione diretta che si basano sulle equazioni:

**Errori, stabilità e condizionamento**

Qualunque metodo per la risoluzione di un sistema lineare produce una soluzione approssimata a causa degli errori di arrotondamento che vengono amplificati e trasmessi. Sia quindi x la soluzione esatta del sistema Ax = b al quale si supponga di applicare un qualunque metodo diretto e sia x + δx la soluzione approssimata che si ottiene. L’influenza dell’algoritmo equivale ad una certa perturbazione δA (matrice di perturbazione), δb (vettore di perturbazione) dei dati iniziali. La soluzione numerica x + δx si può pensare come: (A + δA)(x + δx) = b + δb.

Un algoritmo che produce forti perturbazioni si dice instabile. Si dice stabile un algoritmo che produce perturbazioni modeste. L’entità dell’errore relativo dipende dalla sensibilità della soluzione alle perturbazioni A e b o dal condizionamento del sistema, termine con il quale si designa l’attitudine che ha un dato problema a trasmettere le perturbazioni die dati alla soluzione. Vale il seguente teorema:

nell’ipotesi che la matrice A + δA sia non singolare e che rispetto ad una data norma sia ||δA|| < 1 / ||A-1|| vale la relazione:

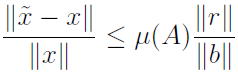
dove μ(A) = ||A|| ||A-1||. Quando μ(A) è “molto grande” si ha una forte amplificazione del membro destro e quindi l’errore sarà molto grande. Per questo μ(A) si dice **numero di malcondizionamento** rispetto alla norma considerata. Il numero è sicuramente non inferiore all’unità perché:

μ(A) = ||A|| ||A-1|| = ||A A-1|| = ||I|| ≥ 1 .

In generale possiamo dire che A si dice malcondizionata se μ(A) >> 1. Per misurare la stabilità o meno dell’algoritmo è anche utile cercare di risalire alla perturbazione ||δA|| . Questa tecnica è detta analisi dell’errore all’indietro. Di validità più generale è invece una maggiorazione a posteriori che si ricava come segue:

sia x\* la soluzione ottenuta per il sistema Ax = b risolto con un qualunque metodo e si abbia b- Ax\* = r. Per r si intende il vettore residuo (in corrispondenza di r = 0 si avrà la soluzione esatta ovvero quella soluzione per cui Ax = b ). Ne segue:

A (x\* - x) = -r quindi (x\* - x) = -A-1 r e per una qualunque coppia di norme matriciali e vettoriali coerenti vale: ||x\*-x|| ≤ ||A-1|| ||r||.

Dalla risoluzione di Ax = b si ricava ||x|| ≥ ||b|| / ||A||.

Mettendo insieme le ultime due disequazioni si ottiene:

**Metodi iterativi in generale**

Molti problemi chiedono la soluzione di un sistema Ax = b di dimensioni molto grandi con matrice A **sparsa** (con pochi elementi non nulli). Se a tale sistema si applica un metodo diretto ci ritroviamo a gestire matrice intermedie **dense** (con molti elementi non nulli) facendo lievitare il costo computazionale. In questi casi conviene quindi usare metodi iterativi in cui ogni iterazione richiede il prodotto di una matrice H (densa quanto A) per un vettore in modo da limitare la mole di calcolo e di memoria utilizzata. Il procedimento generale per costruire un metodo iterativo è il seguente:

dato il sistema Ax – b = 0, con A ε Rn\*n, b ε Cn, det(A) ≠ 0, si trasforma in : x = Hx + c .

Questo si può ottenere in molti modi, ad esempio con una qualsiasi matrice G ε Cn\*n non singolare si può scrivere:

x = x-G(Ax-b) => x = x – Gax + Gb Fattorizzo la x e ottengo: x = x( I-GA ) + Gb che è nella forma che volevo ponendo H = I-GA e c = Gb.

Il processo iterativo quindi è: x(k+1) = Hx(k) + c, con k = 0,1,2,….

Si parte da x(0) che è una approssimazione iniziale della soluzione. La matrice H è detta **matrice di iterazione** e definisce il metodo. Un metodo si dice convergente se la successione x(k) converge alla soluzione del sistema dato.

**Condizione necessaria e sufficiente** affinché un metodo iterativo sia convergente per qualsiasi vettore iniziale x(0) è che la sua matrice di iterazione H sia convergente.

**Dimostrazione:**

sia a la soluzione esatta del sistema Ax = b e si voglia usare un metodo iterativo del tipo x(k+1) = Hx(k) + c con k = 0,1,2,…. . Essendo x = Hx + c equivalente al sistema dato, vale l’identità a = Ha + c quindi sottraendo membro a membro ottengo: x(k+1) – a = Hx(k) – Ha . Indicando con e(k) = x(k) – a l’errore associato a x(k) , quindi si ha che e(k+1) = He(k) . Per ricorrenza abbiamo: e(k+1) = He(k) = H2e(k-1) = … = Hk+1e(0).

Quindi per un arbitrario e(0) si avrà:

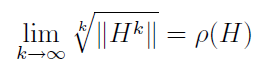
Cioè se la matrice H è convergente.

**Corollario**

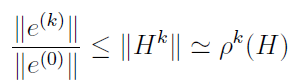
* Per la convergenza del metodo iterativo è necessario e sufficiente che risulti ρ(H) < 1.
* Condizione sufficiente per la convergenza del metodo iterativo è l’esistenza di una norma matriciale per cui si abbia ||H|| < 1.
* Condizione necessaria per la convergenza del metodo iterativo è |det(H)| < 1

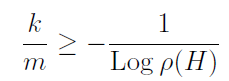
Immagine che contiene tavolo

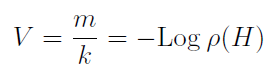
Descrizione generata automaticamente

Infatti si dimostra che:

Quindi asintoticamente (cioè per k abbastanza grande) si ha :

da questa segue che, se ||e(0)|| ≠ 0,

Perciò in un metodo convergente, ||e(k)|| si riduce almeno a ||e(0)||\* 10-m dopo un numero k di iterazioni tale che ρk (H) ≤ 10-m ossia se:

Nell’ambito dei metodi convergenti, la convergenza risulta tanto più rapida quanto più grande è il numero – Log ρ(H). poiché la disequazione k/m è stata dedotta dalle relazioni asintotiche, al numero:

si da il nome di **velocità asintotica di convergenza** del metodo avente matrice di iterazione H.

Se due metodi hanno matrici di iterazione con diverso raggio spettrale è più veloce quello che corrisponde al raggio spettrale minore. Sottraendo ad entrambi i membri il vettore x(k) e tenendo conto che c = -(H-I)a si perviene alla

L’uso di un metodo iterativo comporta l’utilizzo di qualche criterio di arresto. Se ε è una tolleranza d’errore prestabilita, un criterio spesso seguito è il seguente:

tale criterio è inefficiente se il numero ||(H-I)-1|| è molto grande. Per garantire che un metodo iterativo termini dopo un numero massimo N di iterazioni, si affianca ai criteri di arresto scelti l’ulteriore condizione che il calcolo si arresti dopo N iterazioni. Con l’uso dei metodi iterativi non è detto che si annulli l’errore dovuto alle approssimazioni. Possiamo dire però che questo diventa molto contenuto (< ε ).

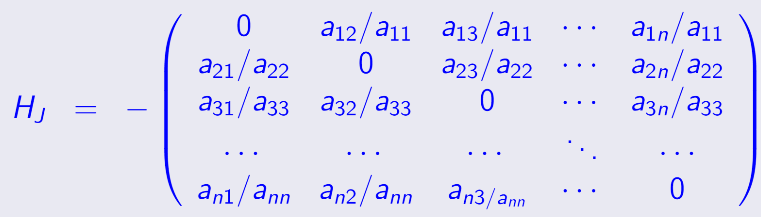
**Metodi di Jacobi**

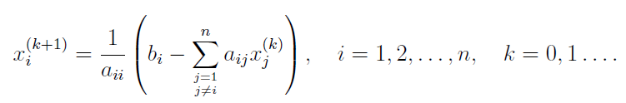
Si supponga A nella forma A = D – E – F dove:

* D è formata dagli elementi diagonali di A.
* E è formata dagli elementi sotto la diagonale di A.
* F è formata dagli elementi sopra la diagonale di A.

Il sistema Ax – b = 0 diventa quindi Dx = (E+F)x + b da cui se aii ≠ 0 si ottiene il metodo di Jacobi:

x(k+1) = D-1(E+F)x(k) + D-1b la cui **matrice di iterazione** è H = D-1(E+F)



Le equazioni di questo sistema sono date da:

Il vettore x(k+1) viene prima memorizzato in una posizione distinta da quella occupata da x(k), poi le n componenti di xi(k+1) vengono trasferite simultaneamente nelle posizioni prima occupate dalle xi(k). Per questo motivo il metodo viene anche chiamato “metodo delle sostituzioni simultanee”.

**Immagine che contiene testo, shoji

Descrizione generata automaticamenteEsempio di costruzione di una matrice di iterazione di Jacobi:**

Immagine che contiene testo, shoji, lavagnabianca

Descrizione generata automaticamente

**Metodo di Gauss-Seidel**

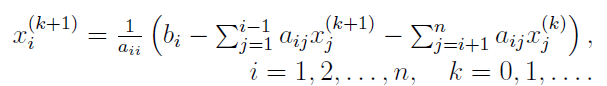
Si supponga A nella forma A = D – E – F dove:

* D è formata dagli elementi diagonali di A.
* E è formata dagli elementi sotto la diagonale di A.
* F è formata dagli elementi sopra la diagonale di A.

Se si scrive il sistema Ax-b= 0 nella forma (D-E)x = Fx + b si ottiene il metodo di Gauss-Seidel.

x(k+1) = (D - E)-1Fx(k) + (D-E)-1b dove la **matrice di iterazione** è data da: HG = (D-E)-1F.

Nel calcolo pratico si fa uso della formulazione:

x(k+1) = D-1Ex(k+1) + D-1Fx(k) + D-1b dove le singole equazioni sono:

Per la matrice di iterazione del metodo di Gauss-Seidel, in generale, non risulta possibile scrivere quali sono i suoi componenti (come nel caso del metodo di Jacobi). Possiamo solo sapere che la prima colonna deve essere nulla e di conseguenza det(HGS) = 0

Immagine che contiene tavolo

Descrizione generata automaticamente**Esempio di costruzione di una matrice di iterazione di Gauss-Seidel:**

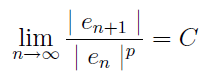
Consente una maggiore economia di memoria rispetto al metodo di Jacobi perché ogni singola componente xi(k+1) appena calcolata può essere subito memorizzata nella posizione prima occupata dalla vecchia componente xi(k). Per questo viene anche chiamato il “metodo delle sostituzioni successive”. Ricordiamo che entrambi i metodi devono soddisfare la condizione aii ≠ 0 (se tale condizione non fosse verificata, è possibile ottenerla ordinando le equazioni e le incognite). Altre analisi che si possono fare sulla matrice A sono le seguenti:

* Se A è una matrice a predominanza diagonale forte, allora i due metodi convergono.
* Se A è una matrice irriducibile a predominanza diagonale debole, i due metodi sono convergenti.
* Se A è una matrice non singolare, i due metodi non convergono.

****In generale la convergenza di uno dei due metodi non implica la convergenza dell’altro.

**Equazioni e sistemi non lineari**

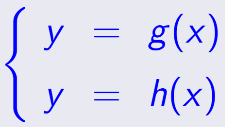
Sia f(x) : A ⊆ R -> R una funzione continua almeno in un intervallo I e si supponga che f(x) non sia della forma f(x) = a1x + a0 con a0,a1 costanti. La relazione f(x) = 0 si dice equazione non lineare nell’incognita x. Il problema di determinare gli zeri (se esistono) raramente può essere risolto con un metodo diretto, cioè effettuando un numero finito di operazioni che portano ad una radice. In generale per il calcolo di una radice α di una equazione non lineare si ricorre a un metodo iterativo, cioè all’applicazione ripetuta di una formula tipo:

dove la funzione φn si dice **funzione di iterazione del metodo** e la sua forma può variare al variare di n. Una opportuna scelta delle funzioni di iterazione e delle approssimazioni iniziali può far convergere la successione{xn} alla radice α cercata. Il calcolo si arresta se si verifica qualche criterio di accettabilità. Ad eccezione dei k valori iniziali, ogni altro termine di {xn} viene calcolato a partire dai k valori iniziali, per questo motivo la formula di sopra viene della “metodo iterativo a k punti”. Se la forma di φn non cambia al variare di n il metodo si dice stazionario. Data una successione {xn} convergente ad un limite α, si ponga en = xn – α; se esistono due numeri reali p ≥ 1 e C ≠ 0 tali che sia;

si dice che la successione ha ordine di convergenza p e fattore di convergenza C. per p = 1 e p = 2 si dice anche lineare e quadratica rispettivamente. Nel caso di p = 1, la convergenza ad α implica C < 1. Analogamente a quanto stabilito per i sistemi lineari, anche nei sistemi non lineari si può esprimere una velocità asintotica di convergenza:

* Nel caso p = 1 , C < 1 si pone V = -log ( C ) .
* Nel caso p > 1, la velocità ha un’espressione più complicata che dipende dai valori di p e di C e cresce al crescere di p.

**Metodo di separazione grafica**

Si pone f(x) = g(x) – h(x) dove le funzioni g e h possono essere scelte in tanti modi. Se vale l’uguaglianza f(x)=0 allora deve valere anche g(x) = h(x). Questa equazione porta (introducendo la variabile y) alla scrittura del sistema:

Le soluzioni (se esistono) sono coppie ordinate di punti (x,y) le cui componenti x coincidono con le soluzioni dell’equazione f(x) = 0. La tecnica di separazione grafica consiste nel tracciare il grafico delle funzioni g(x) e h(x) e controllare i loro punti di intersezione (più precisamente le coordinate). Dallo studio dei due grafici è possibile dire quante sono le soluzioni di f(x) = 0 ed individuare gli intervalli di separazione a ciascuno dei quali appartiene una ed una sola soluzione dell’equazione.

**Esempio**

**Immagine che contiene aquilone, testo, colorato

Descrizione generata automaticamente**Si vuole risolvere l’equazione: 5x2 – 2ex = 0. La separazione più ovvia è quella di porre g(x) = 5x2 e h(x) = 2ex. Tracciando il grafico delle due funzioni otteniamo

Dalla figura si deduce che l’equazione data ha 3 soluzioni:

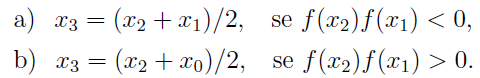
* α1 ε [-1,0]
* α2 ε [0.5 , 1.5]
* α3 ε [3, 3.5]

La separazione grafica risulta un metodo valido ma non infallibile. Il passo successivo consiste nell’applicare un metodo di approssimazione delle soluzioni.

**Metodo di bisezione**

Questo è un metodo iterativo a due punti non stazionario. È il più semplice metodo iterativo per approssimare gli zeri reali di una funzione f(x). Si costruisce un intervallo contenente uno zero di f(x) e si assume come approssimazione di tale zero, l’ascissa del punto medio dell’intervallo. Sia f(x) continua in un intervallo [a,b] e poniamo x0 = a, x1 = b. Supposto che si abbia f(x0)(x1) < 0 , per la continuità si avrà almeno uno zero nell’intervallo [x0 , x1] (per semplicità supponiamo che ne abbia esattamente uno ). Il numero x2 = (x0 + x1 ) / 2 sarà certamente un’approssimazione migliore di α rispetto ad almeno una delle precedenti x0, x1. Se non si verifica che f(x2) = 0 si confronta il segno di f(x2) con il segno di f(x1):

* se risulta f(x1)(x2) < 0 allora α ε [x2,x1] e si ripete il procedimento.
* se risulta f(x1)(x2) > 0 allora α ε [x0,x2] e si ripete il procedimento.

 Quindi la nuova approssimazione x3 sarà data da:

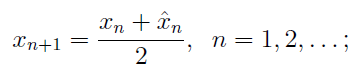
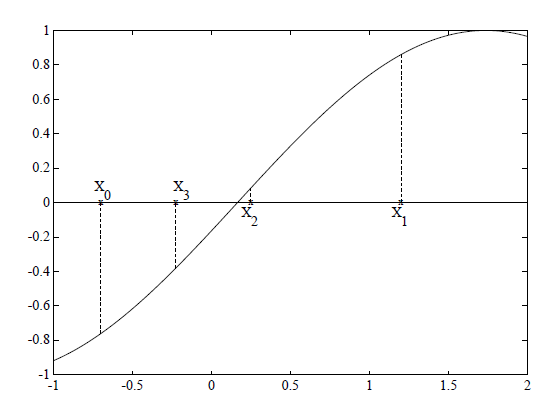
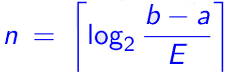
Ripetendo il procedimento si determinano x4, x5, x6 … secondo la formula generale:

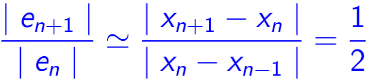
Immagine che contiene testo

Descrizione generata automaticamenteDove per n > 1 si ottiene:

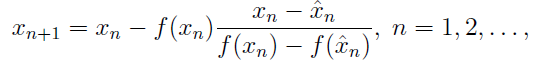
Poiché ad ogni passo l’intervallo contenente α viene dimezzato, dopo n passi si ha un’approssimazione xn+1 tale che:

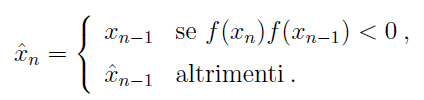
Da cui segue che il limn->∞ |xn – α | = 0 , che prova la convergenza del metodo alla radice α. Possiamo usare come possibile criterio di arresto la maggiorazione dell’errore assoluto: ½n (b-a) < ε, dove ε > 0 è un numero opportunatamente prefissato. Per ottenere l’approssimazione richiesta bastano n iterazioni con n =

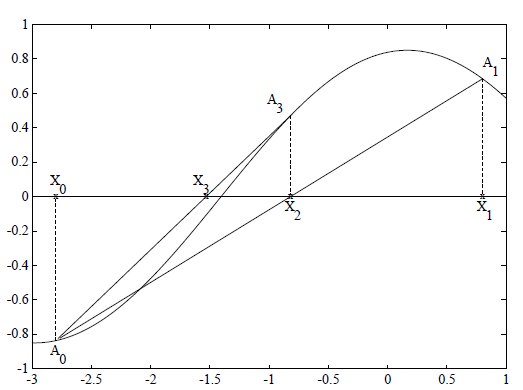
Il metodo converge linearmente ma è molto lento quindi viene usato per trovare una prima approssimazione, infatti, assumendo |xn – xn-1| come stima di |xn – α | = |en|, si ha per n abbastanza grande, l’uguaglianza approssimata



**Regula falsi e metodo delle secanti**

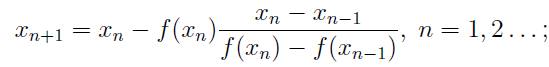
Si abbiano le stesse condizioni iniziali del metodo di bisezione. Il metodo di regula falsi (metodo di falsa posizione) è un altro metodo iterativo stazionario a due punti nel quale, partendo da un intervallo ]x0, x1[ , ad ogni passo si costruisce un nuovo intervallo di estremi xn e xn\* contenente uno zero di f(x) e si assume come approssimazione xn+1 lo zero di una funzione lineare il cui grafico è la retta passante per i punti An = [xn , f(xn)] e An\* = [xn\*, f(xn\*)]. Si ha quindi lo schema:

* Per n = 1 si pone x1\* = x0 .
* Per n > 1 si pone, come nel metodo di bisezione:

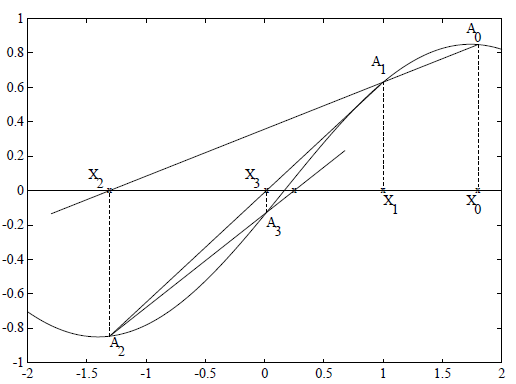


Se f(x) ε C2 (I) e se il metodo regula falsi converge ad uno zero α di f(x) con α ε I tale che f’(α)̠ ≠ 0 e f’’(α)≠0, la convergenza è di ordine p = 1.618…

**Metodo delle secanti**

È una variante del metodo regula falsi. È un metodo iterativo stazionario a due punti. Sono richieste solo due approssimazioni iniziali senza alcun’altra condizione e senza la necessità di controllare il segno di f(x) ad ogni passo. In questo metodo il calcolo dell’approssimazione xn+1 utilizza le informazioni precedenti secondo la formula:

Il numero xn+1 è individuato geometricamente dal punto in cui la secante passante per i punti An = [xn, f(xn)] e An-1 = [xn-1, f(xn-1)] incontra l’asse delle ascisse. Per n ≥ 2, ad ogni passo il calcolo di xn+1 richiede la sola valutazione di f(xn). la convergenza del metodo è garantita se le approssimazioni iniziali e x0 e x1 sono scelte abbastanza vicine ad α.



Se f(x) ε C2 (I) e se il metodo delle secanti converge ad uno zero α di f(x) con α ε I tale che f’(α)̠ ≠ 0 e f’’(α)≠0, la convergenza è di ordine p = (1+√5) ~ 1.618

**Metodi iterativi a un punto**

Data l’equazione f(x) = 0, si può sempre costruire una funzione φ(x) tale che x = φ(x). Per esempio, basta porre φ(x) = x – g(x)f(x) dove g(x) è una funzione continua che si suppone diversa da 0 nei punti che appartengono ad una regione contenente gli 0 di f(x). Se α è uno 0 di f(x) (ovvero f(α) = 0) si ha che α = φ(α) e viceversa. Allora α si dice **punto fisso** della trasformazione x= φ(x) ( o più semplicemente di φ(x) ). Per ogni scelta della funzione φ(x) si può considerare un metodo iterativo stazionario ad un punto della forma xn+1 = φ(xn) e il problema di approssimare uno zero α di f(x) si riduce a costruire una successione convergente ad α, punto fisso di φ(x).

**Condizione sufficiente per la convergenza locale (Teorema di convergenza locale)**

Analizziamo una condizione sufficiente per la convergenza dipendente dalla scelta di x0. Sia α = φ(α), α ε I, φ ε C1(I). Se esistono due numeri ρ e K con K < 1 tali che per ogni x ε [α – ρ, α + ρ] C I si verifichi la condizione: |φ’(x)| ≤ K, allora per il metodo xn+1 = φ(xn), n = 0,1,2,.. valgono le seguenti proposizioni:

1. Se x0 ε ]α–ρ, α+ρ[ => xn ε ]α-ρ , α+ρ[ per n ε N.
2. Se x0 ε ]α–ρ, α+ρ[ => limn->∞ xn = α
3. α è l’unico punto fisso di φ(x) in [α-ρ,α+ρ].

**Dimostrazione del punto 1**

Si dimostra per induzione, cioè, scelto un x0 ε ]α–ρ, α+ρ[ si ammette per ipotesi che sia, per un certo n, che sia xn ε ]α–ρ, α+ρ[ ossia |xn – α| < ρ e si deduce che deve essere xn ε ]α–ρ, α+ρ[ ovvero |xn+1 – α| < ρ. Infatti facendo uso della struttura del metodo e del teorema del valor medio, si ha:

xn+1 – α = φ(xn) – φ(α) = φ’(ξn)(xn – α)

Dove ξn è compreso tra xn e α. Dall’ipotesi fatta su xn e da quelle del teorema segue:

|xn+1 – α| = φ’(ξn)||xn – α| ≤ k|xn – α| < ρ

**Dimostrazione del punto 2**

La proposizione 2 segue dall’ipotesi 0 < K < 1 e dalla disuguaglianza |xn+1 – α| ≤ Kn+1|x0 – α| che si ottiene per ricorrenza dall’ultima relazione della dimostrazione del punto 1.

**Dimostrazione del punto 3**

La proposizione 3 si dimostra per assurdo. Se in ]α-ρ, α+ρ[ esistesse un altro punto β≠α si avrebbe |α-β| = |φ(α) – φ(β)| = |φ’(ξ)||α-β| ≤ K|α-β| < |α-β|.

In generale l’ordine di convergenza di un metodo iterativo è un numero reale p ≥ 1. Per i metodi iterativi stazionari ad un punto vale il seguente teorema:

**Teorema sull’ordine di convergenza**

Un metodo iterativo stazionario ad un punto, la cui funzione di iterazione φ(x) sia sufficientemente derivabile, ha ordine di convergenza uguale ad un numero intero positivo p. Se il metodo converge ad α, la convergenza è di ordine p allora e solo che si abbia:

φ(α) = α, φ(i)(α) = 0, per 1 ≤ i < p , φ(p)(α) ≠ 0

**Dimostrazione**

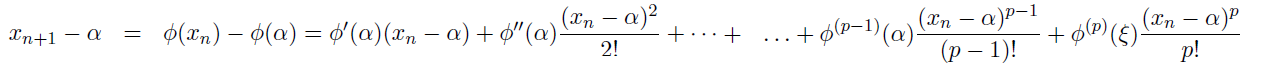
Usando la formula di Taylor si ha:

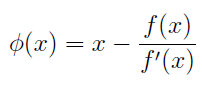
Immagine che contiene testo, orologio

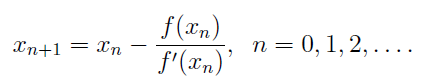
Descrizione generata automaticamentedove ξ è compreso tra xn e α. Se valgono le ipotesi sulle derivate successive di φ calcolate in α, si ha

da cui si deduce che l’ordine di convergenza è p ed il fattore di convergenza vale C = |φ(p)(α)|/ p!.

Viceversa se l’ordine è p, sia φ(i)(α) la prima derivata non nulla nel precedente sviluppo di Taylor intorno al punto α. Se fosse i ≠ p , per il ragionamento diretto anche l’ordine sarebbe i ≠ p contro l’ipotesi per cui deve risultare i = p e quindi valgono le ipotesi sulle derivate di φ. Come criterio di arresto si può assumere la condizione: |xn+1 – xn| ≤ ε.

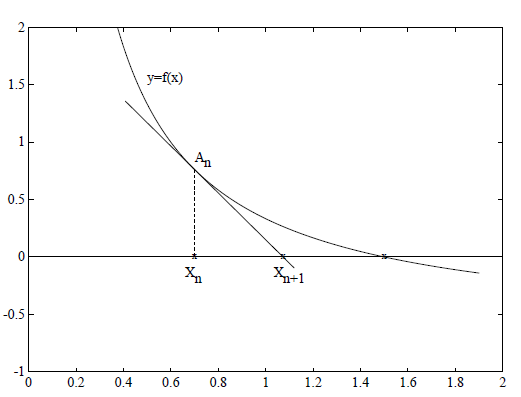
**Metodo di Newton (Metodo delle tangenti)**

Si può applicare per approssimare uno zero α di f(x) se in tutto un intorno di α, f(x) è derivabile con continuità. In tal caso, assumendo la funzione di iterazione della forma



Si ha il metodo:

L’iterata xn+1 è individuata dal punto d’incontro dell’asse delle ascisse con la tangente alla curva y = f(x) nel punto An = [xn, f(xn)] per questo si chiama anche metodo delle tangenti.



Sulla convergenza e l’ordine del metodo di Newton, vale il seguente **teorema**:

sia f(x)ε C3 ([a,b] ), a < α < b, f(α) = 0, f’(α)≠0, f’’(α)≠0, allora valgono le 3 proposizioni:

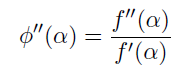
1. La convergenza è di ordine p ≥ 2.
2. Se p = 2, il fattore di convergenza è:

Immagine che contiene testo, orologio

Descrizione generata automaticamente**Dimostrazione del punto 1**

Per il primo punto, dalla struttura di φ(x) si ricava:

Da cui segue che φ’(α) = 0. Questo risultato ci dice che per zeri semplici, il metodo di newton ha ordine di convergenza almeno 2.

**Dimostrazione del punto 2**

Per dimostrare l’asserto 2 procedo con il calcolo delle derivate e ottengo:

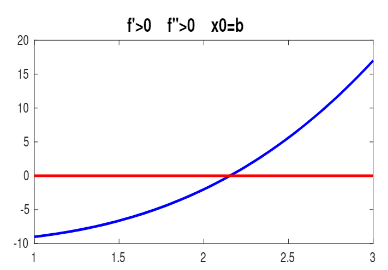
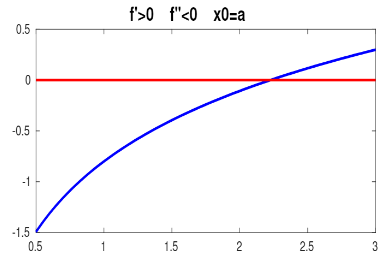
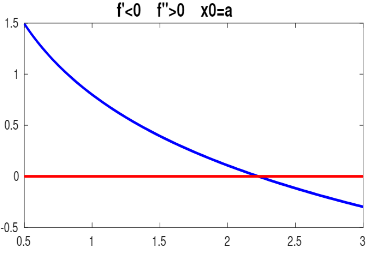
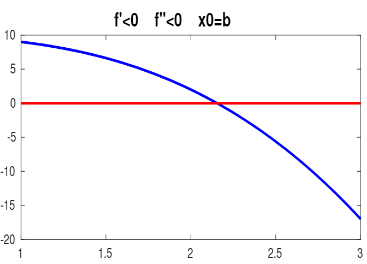
Segue che se f’’(α) ≠ 0 allora anche φ’’(α)≠ 0. Questo risultato garantisce che l’ordine di convergenza è p=0. Se invece f’’(α) = 0 è anche φ’’(α) = 0 si ha p > 2.

La seconda tesi del teorema discende direttamente dal teorema generale sull’ordine di convergenza con p = 2.

Si deve osservare che la convergenza di Newton è di tipo locale, cioè si verifica se si sceglie x0 sufficientemente vicino ad α. A tale punto iniziale si può arrivare per esempio, con il metodo di bisezione. Nel caso di radici semplici, per il metodo di Newton esistono sicuramente punti iniziali x0 che lo rendono convergente. Infatti, risultando φ’(α) = 0, per la continuità della funzione, esiste un intorno di α per il quale vale la condizione principale del teorema di convergenza locale.

**Condizioni di convergenza**

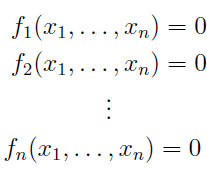
Data l’equazione f(x) = 0, sia α una soluzione di molteplicità dispari con [a,b] intervallo di separazione per α ed inoltre sia f ε C2([a,b]). Se f’(x) e f’’(x) sono di segno costante in [a,b] allora in metodo di Newton converge scegliendo come punto iniziale un valore x0 ε [a,b] tale che f(x0)f’’(x0) > 0. Per avere la costanza del segno delle prime due derivate di f(x) si hanno 4 casi possibili:

1. f’(x) > 0 ; f’’(x) > 0
2. f’(x) > 0 ; f’’(x) < 0
3. f’(x) < 0 ; f’’(x) > 0
4. f’(x) < 0 ; f’’(x) < 0

**Immagine che contiene tavolo

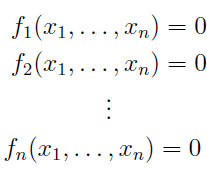
Descrizione generata automaticamente**

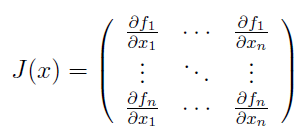
**Metodi iterativi in Rn**

La teoria dei metodi iterativi precedentemente esposta si può estendere al caso in cui sia f(x):Rn->Rn , cioè al caso di un sistema di n equazioni (non lineari) in n incognite.

A questo punto la soluzione α da approssimare è della forma: α = (α1,α2,…,αn)T come punto fisso di una opportuna funzione di iterazione φ(x):Rn->R data da φ(x) = x – G(x)f(x) con G che è una matrice n\*n non singolare in un dominio D ⊆ Rn contenente α. Si considerano quindi metodi iterativi della forma:

**Metodo di Newton-Raphson**

Per estendere il metodo di Newton al sistema non lineare:

supponiamo che le funzioni fi siano derivabili con continuità rispetto a ciascuna variabile e che la matrice jacobiana

sia non singolare in un dominio contenente nel suo interno una soluzione α del sistema. Specializzando la funzione di iterazione nella forma: φ(x) = x – J-1(x)f(x) si ha il sistema x = x – J-1(x)f(x) da cui il metodo iterativo di Newton o di Newton-Raphson

Nell’uso pratico l’iterata x(k+1) si trova dalla soluzione del sistema lineare

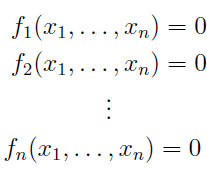
Dove d(k) = x(k+1) – x(k) .

Nel caso di una lenta variazione della matrice Jacobiana J(x), si può ricorrere ad un metodo di Newton semplificato della forma:

J(x(0))d(k) = -f(x(k))

dove la matrice jacobiana è valutata una sola volta in una buona approssimazione iniziale x(0). Se il metodo di Newton semplificato converge, la convergenza è in genere lineare.

Sull’ordine di convergenza del metodo si ha il seguente teorema:

Sia α ε D una soluzione del sistema

e le funzioni fi siano della classe C3(D) e tali che J(x) sia non singolare in D; allora l’ordine di convergenza del metodo di Newton è almeno p = 2 e si ha convergenza per ogni scelta di x(0) in un opportuno dominio contenente α.

**Zeri di polinomi**

L’equazione x = φ(x) sia algebrica di grado m ≥ 2, cioè della forma

con i coefficienti ai reali. All’equazione sono applicabili tutti i metodi visti nei paragrafi precedenti ma le particolari proprietà della classe dei polinomi forniscono ulteriori procedimenti per la localizzazione e l’approssimazione delle radici. Uno strumento utile in tale senso è la successione di Sturm:

si dice **Successione di Sturm** per il polinomio P(x) con ai reali, una successione di polinomi reali

per cui valgono le seguenti proprietà:

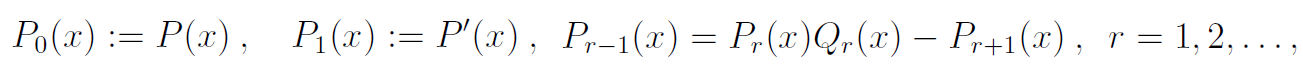
* Pk(x) non ha zeri reali.
* α ε R, Pr(α) = 0 per i ≤ r ≤ k-1 implica Pr-1(α)Pr+1(α) < 0 .
* α ε R, P0(α) = 0 implica P0’(α)P1(α) > 0. Da questa proprietà segue che gli zeri reali di P(x) sono tutti semplici.

**Teorema di Sturm**

Se la successione

è una successione di Sturm il numero degli zeri del polinomio P(x) nell’intervallo a < x ≤ b è dato da V(a) – V(b) dove V(x) è il numero delle variazioni di segno presenti nella successione dei valori non nulli assunti dal polinomio di sopra nel punto x.

**Corollario**

Se i polinomi della successione di Sturm sono m+1 (quindi m=k) la successione si dice completa e se tutti i polinomi hanno i coefficienti di grado massimo dello stesso segno, allora l’equazione P(x) = 0 ha m radici reali e distinte e viceversa. Per costruire effettivamente una successione di Sturm per il polinomio P(x), si pone

dove Qr(x) e -Pr+1(x) sono rispettivamente quoziente e resto della divisione Pr-1(x) / Pr(x) . Il processo si arresta poiché il grado dei polinomi decresce al crescere dell’indice e perciò per un certo k ≤ m risulta Pk-1(x) =Pk(x)Qk(x) . Il processo di costruzione della successione di Sturm è riconducibile al noto algoritmo di Euclide che fornisce il massimo comune divisione di P0(x) e P1(x) cioè si ha:

Segue che, nel caso in cui P(x) e P’(x) non abbiano zeri reali in comune, i polinomi P0(x), P1(x),…,Pk(x) formano una successione di Sturm per P(x).

**Costruzione della successione di Sturm**

Immagine che contiene testo

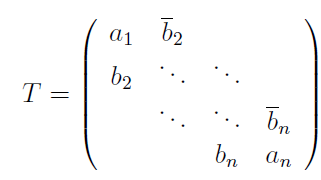
Descrizione generata automaticamentePk(x) non ha zeri reali e quindi la proprietà 1 risulta verificata. Se Pr(α) = 0 con 1 ≤ r ≤ k-1, allora si ha Pr-1(α) = -Pr+1(α) ≠ 0 e quindi è verificata la proprietà 2: Pr-1(α)Pr+1(α) < 0. Dalla definizione di P0(x) e P1(x), se P(α) = 0 segue la proprietà 3: P0’(α)P1(α) = (P’(α))2 > 0 . Se P(x) e P’(x) hanno zeri in comune, questi sono tutti zeri di Pk(x) e la successione ottenuta con l’algoritmo di Euclide non fornisce una successione di Sturm. In tal caso si può verificare che la successione:

è una successione di Sturm per il polinomio P(x) / Pk(x) che ha tanti zeri semplici quanti sono gli zeri distinti di P(x).

Immagine che contiene orologio

Descrizione generata automaticamenteLa successione di Sturm può essere usata per individuare un intervallo [a,b] contenente una sola radice reale α di una equazione algebrica e quindi, con successivi dimezzamenti dell’intervallo come nel metodo di bisezione, si può approssimare α con qualsiasi accuratezza. Il particolare, per approssimare α si può usare il metodo di Newton:

**Una particolare successione di Sturm**

Si considera la matrice tridiagonale hermitiana:

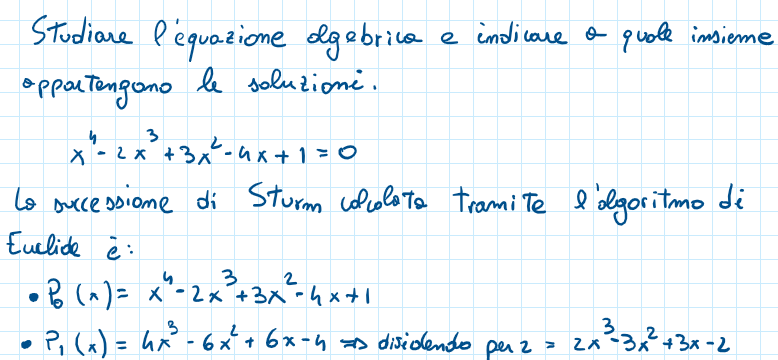
con bi ≠ 0. Il suo polinomio caratteristico può essere calcolato per ricorrenza ottenendo P(λ) = Pn(λ) dove

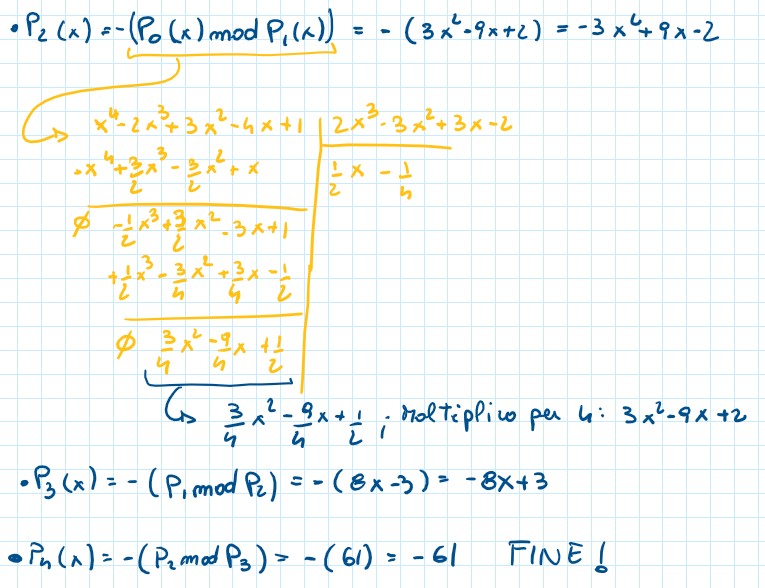
con P0(λ) = 1 e P1(λ) = a1 – λ.

**Teorema**

Nelle ipotesi fatte su T, la successione

è una successione di Sturm per P(λ) e gli zeri di P(λ) sono distinti.

**Esempio**

****

**Immagine che contiene tavolo

Descrizione generata automaticamente**

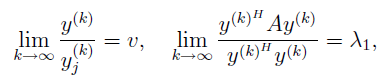
**Calcolo di autovalori ed autovettori**

Per stimare gli autovalori e gli autovettori di una matrice A sembrerebbe naturale ricorrere all’approssimazione delle radici dell’equazione caratteristica det(A-λI) = 0 e successivamente, per ogni autovalore λ trovato, risolvere il sistema lineare omogeneo (A-λI)x = 0. Tuttavia è sconsigliabile seguire tale via tranne per casi particolari.

**Metodo delle potenze (da sapere bene)**

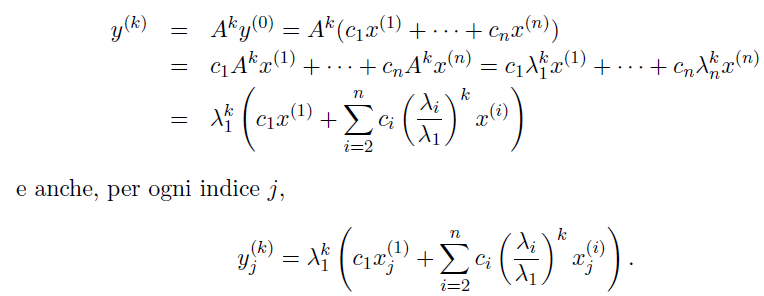
Serve ad approssimare un autovalore di modulo dominante ed un autovettore ad esso associato. Si basa sul teorema:

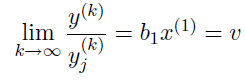
Sia A ε Cn\*n una matrice diagonalizzabile con autovalori soddisfacenti la condizione |λ1| > |λ2|≥|λ3,…,n|, sia z(0) ε Cn un vettore arbitrario non nullo, allora il processo iterativo:

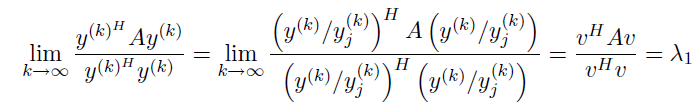
y(0) = z(0)  ; y(k) = Ay(k-1) è tale che:

dove j è un indice per cui yj(k) ≠ 0 e v è l’autovettore associato a λ1.

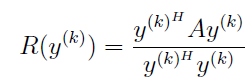
**Dimostrazione**

La diagonalizzabilità di A implica l’esistenza di n autovettori x(i), i = 1,2,…,n linearmente indipendenti e quindi che il vettore z(0) possa rappresentarsi nella forma z(0) = Σcix(i) , dove è lecito pensare che sia c1 ≠ 0. Da y(k) = Ay(k-1) segue:

Scegliendo yj(k) ≠ 0, tenendo conto dell’ipotesi dell’autovalore di modulo dominante e dividendo membro a membro si ottiene:

Dove b1 = 1/xj(1); perciò il vettore v è l’autovettore associato a λ1, come afferma la tesi. Si ha quindi che Av = λ1v e anche vHAv = λ1vHv da cui λ1 = (vHAv)/(vHv).

Inoltre



Il rapporto si dice **Quoziente di Rayleigh.**

In pratica si adotta come criterio di arresto la condizione |R(y(k)) – R(y(k-1)| < ε con ε > 0 prefissato.

L’ipotesi che A sia diagonalizzabile è in genere difficile da verificare ma questa è sicuramente vera per la vasta classe delle matrici normali AHA = AAH (matrici normali). Il teorema si può estendere al caso più generale in cui λ1 abbia molteplicità r ≥ 1 modificando l’ipotesi |λ1| > |λ2|≥|λ3,…,n| nella forma λ1 = λ2 = … = λr , |λ1| > |λr+1| ≥ … ≥ |λn| . Si osservi che per r = 1 l’unico vettore associato a λ1 è approssimato da z(k) mentre per r > 1, z(k) approssima un autovettore appartenente allo spazio generato dagli altri r autovettori associati a λ1 e l’autovettore approssimato cambia in dipendenza dalla scelta del vettore iniziale z(0) . Nell’ipotesi che |λ1| > |λ2| > |λ3| ≥ … |λn|, il metodo delle potenze consente il calcolo dell’autovalore λ2 e del corrispondente autovettore utilizzando la conoscenza di λ1 e x1.

Immagine che contiene testo

Descrizione generata automaticamenteConsideriamo la matrice A1 = A – λ1x1x1H . A1x(1)  = 0 e A1x(i) = λix(i) . La matrice A1 ha autovalori 0, λ2,λ3 …, λn e gli stessi autovettori di A. Il metodo applicato ad A1 converge all’autovettore x(2) e all’autovalore λ2. Questo procedimento si chiama **deflazione** e può essere ripetuto fin tanto c’è un autovalore di modulo massimo.

**Immagine che contiene testo

Descrizione generata automaticamenteEsempio:**

**Il metodo di Jacobi per matrici simmetriche**

Il metodo di Jacobi permette di approssimare tutti gli autovalori di una matrice hermitiana (simmetrica nei reali). Per semplicità si considerano solo matrici A reali e simmetriche. Il metodo consiste nell’operare successive trasformazioni per similitudine mediante l’uso di matrici di rotazione Grs della forma:

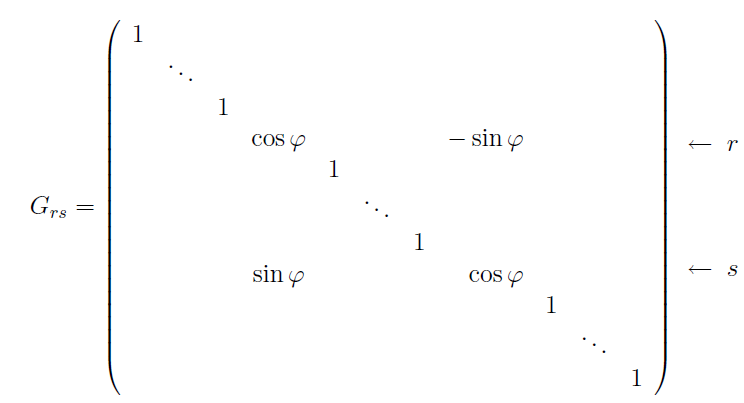


Immagine che contiene orologio

Descrizione generata automaticamenteTali matrici sono determinate dagli interi r ed s, con 1 ≤ r < s ≤ n e dal parametro φ, inoltre sono ortogonali, cioè si ha

Immagine che contiene testo

Descrizione generata automaticamenteIl metodo di Jacobi è un metodo iterativo in cui al passo k-esimo si trova una matrice Ak mediante una matrice ortogonale Grs(k) secondo l’algoritmo seguente:

Gli indici r ed s variano al variare di k. Nella versione classica del metodo si scelgono gli indici di Grs(k) coincidenti con quelli di un elemento non diagonale ars(k) di Ak avente modulo massimo e il valore di φ viene determinato in modo che nella matrice trasformata Ak+1 risulti ars(k+1) = 0 .

La trasformazione Ak+1 = Grs(k)TAkGrt(k) lascia inalterati tutti gli elemeni di Ak non appartenenti alle righe e alle colonne di indici r ed s, cioè si ha:

Immagine che contiene testo

Descrizione generata automaticamenteVengono modificati solo gli elementi delle righe e delle colonne di indici r ed s e inoltre si conserva la simmetria iniziale.Precisamente si ha:

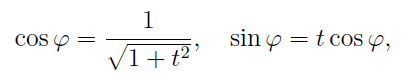
Dalla prima è verificata per ogni φ soluzione dell’equazione:



In pratica si ricavano direttamente due valori, sinφ, cosφ che servono a costruire la matrice Grs(k), scrivendo la precedente equazione nella forma:

Immagine che contiene testo, orologio

Descrizione generata automaticamenteDove si è posto: t = tanφ, m = (arr(k) – ass(k) ) / 2ars (k) . Scegliendo fra le due radici quella di modulo minore, corrispondente a |φ| ≤ π/4 si ha:

Per m = 0 si sceglie t = 1, in ogni caso si ottiene:

Con cui si costruisce la Grs(k) cercata. In questo modo per ogni passo si esclude un elemento non diagonale e il suo simmetrico. Vale il seguente teorema:

**Teorema di Jacobi**

Immagine che contiene testo, orologio, antenna, calibro

Descrizione generata automaticamentela successione {Ak} generata con il metodo di Jacobi classico converge alla matrice diagonale D = diag(λ1, λ2,…,λn) dove λi sono gli autovalori della matrice A. In base a questo teorema, un criterio di arresto delle iterazioni potrebbe essere il verificarsi della condizione

con ε > 0 prefissato.

**Schema del metodo QR**

L’approssimazione simultanea di tutti gli autovalori di una matrice quadrata qualunque, viene generalmente effettuata mediante il metodo QR.

**Teorema (fattorizzazione QR)**

Per ogni matrice A ε Rn\*n esiste una decomposizione nel prodotto di una matrice Q ortogonale per una matrice R tridiagonale superiore.

**Immagine che contiene orologio, calibro

Descrizione generata automaticamente**La dimostrazione di questo teorema si ottiene costruendo effettivamente due matrici Q ed R che si può fare premoltiplicando successivamente A con matrici di rotazione Grs scelte in modo che ad ogni passo si ottenga una matrice prodotto in cui risulti nullo un elemento di indici (s,r) situato sotto la diagonale principale (nel caso in cui fosse già nullo si pone Grs = I ). La strategia è quella di premoltiplicare A ordinatamente per le n(n-1)/2 matrici G12, G13, G14….,Gn-1,n in modo che, nelle successive matrici prodotto, risultino nulli rispettivamente gli elementi di indici (21),(31)….,(n,n-1). Poiché gli elementi che si annulla stanno tutti sotto la diagonale principale, si otterrà alla fine una matrice triangolare superiore e si ha che Gn-1,n \*… \* G13\*G12 = R . Essendo le Grs ortogonali si ha A = G12T G13T … Gn-1,nT \* R quindi A = QR con Q = G12T G13T … Gn-1,nT . Il teorema si può generalizzare al caso di una matrice ad elementi complessi. In tal caso Q della fattorizzazione è una matrice unitaria. L’algoritmo QR per la ricerca degli autovalori di A ε Rn\*n è la seguente:

Se la matrice A è della forma di Hessenberg superiore oppure tridiagonale, la fattorizzazione QR richiede al più n-1 premolitiplicazioni per matrici di rotazione, essendo al più n-1 gli elementi non nulli di A al disotto della diagonale principale. La giustificazione teorica del metodo QR si completa con i due teoremi seguenti che ci si limita ad enunciare nel caso reale:

Immagine che contiene testo

Descrizione generata automaticamente**Teorema di Schur:** Per ogni matrice A ε Rn\*n esiste una matrice reale ortogonale B tale che la trasformata per similitudine S = B-1AB = BTAB è una matrice triangolare a blocchi con i blocchi diagonali di ordine uno due, della forma:

I blocchi diagonali di ordine 1 sono autovalori reali di A, mentre ogni blocco diagonale di ordine due ha come autovalori una coppia di autovalori complessi e coniugati di A.

Immagine che contiene testo

Descrizione generata automaticamente**Teorema**: se gli autovalori di A ε Rn\*n sono reali e distinti in modulo e quindi verificano la condizione: |λ1| > |λ2| > … |λ3| e se gli autovalori di A formano una matrice X tale che X-1 sia fattorizzabile LR, allora le matrici Ak per k->∞ tendono ad una matrice triangolare superiore e gli elementi diagonali aii(k) di Ak tendono agli autovalori λi di A ordinati per modulo decrescente. Se A possiede qualche coppia di autovalori complessi e coniugati ma i moduli di ciascuna coppia e quelli degli autovalori reali sono distinti, allora le matrici Ak convergono ad una matrice triangolare a blocchi del tipo

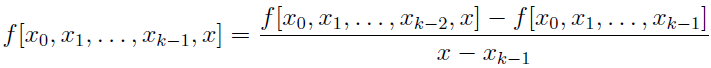
In cui gli autovalori dei vari blocchi sono ordinati per modulo decrescente.

**Interpolazione e approssimazione**

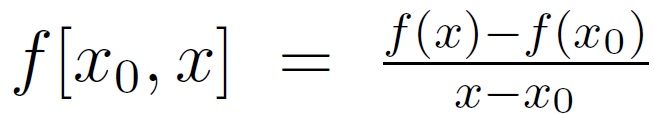
In molti casi si ha a che fare con una funzione f(x) di forma sconosciuta di cui si possiede solo una tabulazione in un numero finito di punti. In questi casi è possibile stimare la funzione f(x) in un punto sconosciuto mediante i dati conosciuti. Questa operazione è detta **interpolazione** e solitamente si effettua sostituendo f(x) con una funzione facile da calcolare come un polinomio.

**Differenze divise**

Per poter definire il polinomio di interpolazione nella forma di Newton dobbiamo introdurre le differenze divise. Assegnata una funzione f(x) : I ⊆ R -> R si ha la seguente definizione:

siano x0,x1,…,xk-1 ε I con xi ≠ xj per i ≠ j, la funzione:

dove per k = 1



è definita per ogni x ε I, x ≠ xi con i = 0,1,…,k-1e si chiama **differenza divisa di grado k.** Se f(x) è derivabile su I e la funzione f[x0,x1,---,xk-1,x] si estende a tutto I, vale il seguente teorema:

Immagine che contiene tavolo

Descrizione generata automaticamente

Quadro delle differenze divise

Il quadro può essere esteso quanto si vuole. Nel caso in cui gli elementi della colonna delle differenze divise di ordine r risultino tutti uguali fra loro, gli elementi delle colonne successive sono nulli, perciò il grado del polinomio di interpolazione è proprio r.

**Teorema di espansione**

Sia f(x) : I ⊆ R -> R e siano x0,x1,x2, … ,xk ε I con xi ≠ xj se i ≠ j, allora vale l’identità:

**Immagine che contiene testo

Descrizione generata automaticamente**

**Proprietà delle differenze divise**

* **Simmetria**

Se i0,i1,…,ik è una qualunque permutazione dei numeri 0,1,…,k si ha:

* **Prolungamento per continuità**

Se f(x) ε C1(I), la funzione f[x0,x1,…,xk-1,x] è prolungabile per continuità su tutto I.

**Interpolazione parabolica**

Immagine che contiene testo, orologio, calibro, dispositivo

Descrizione generata automaticamenteSiano dati k+1 punti reali x0,x1,…,xk ε I, due a due distinti, in corrispondenza dei quali sono noti i k+1 valori reali f(x0),f(x1),…,f(xk). L’interpolazione parabolica consiste nel determinare un polinomio di grado al più k:

tale che:

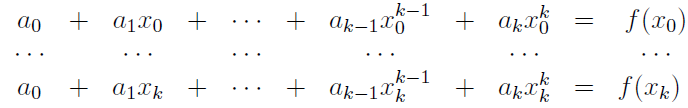
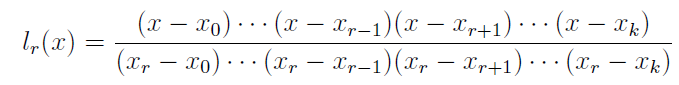
il polinomio Pk(x) si chiama **polinomio di interpolazione**. Nell’insieme dei polinomi ne esiste uno solo che verifica Pk(xi) = f(xi) per ogni i, infatti imponendo il sistema lineare di k+1 equazione nelle k+1 incognite ai si ottiene:

Immagine che contiene testo, stormo

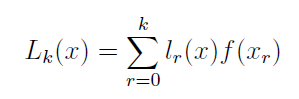
Descrizione generata automaticamenteIl sistema ha la seguente matrice dei coefficienti detta matrice di Vandermonde:

Il determinante della matrice di Vandermonde è : e risulta ≠ 0 essendo i punti xi a due a due distinti. Il polinomio di interpolazione può risultare di grado minore a k se nella soluzione del sistema, risulta ak = 0 per questo si è soliti dire che con k+1 punti il polinomio di interpolazione risulta di grado al più k. Il sistema lineare Va = f (dove a = (ao,a1,…,ak)T e f ( f(x0), f(x1),f(x2),…f(xk) ) ha un’unica soluzione e perciò il polinomio di interpolazione risulta unico.

**Polinomio di interpolazione di Lagrange**

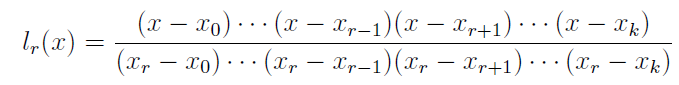
Una forma di polinomio di interpolazione si può ottenere per mezzo delle funzioni polinomiali di grado k:

r = 0,1,…,k. I polinomi di questo tipo godono della proprietà

di conseguenza il polinomio



verifica le condizioni

Lk(x) si chiama polinomio di interpolazione di **Lagrange** e i polinomi

Sono detti polinomi fondamentali della interpolazione di Lagrange.Immagine che contiene testo, orologio, calibro, dispositivo

Descrizione generata automaticamente L’errore che si commette se si sostituisce alla funzione f(x) il polinomio di interpolazione Pk(x) o Lk(x) si ricava da

**Immagine che contiene tavolo

Descrizione generata automaticamenteEsempio polinomio di interpolazione di Lagrange**

Immagine che contiene tavolo

Descrizione generata automaticamente

**Polinomio di interpolazione di Newton (Teorema)**

Un altro polinomio di interpolazione è della forma:

noto come polinomio di interpolazione di Newton.

**Immagine che contiene tavolo

Descrizione generata automaticamenteEsempio 1**

**Esempio 2**

**Immagine che contiene tavolo

Descrizione generata automaticamente**

**Errore nell’interpolazione parabolica**

L’errore che si commette se si sostituisce alla funzione f(x) il polinomio di interpolazione Pk(x) o Lk(x) (il polinomio di interpolazione è unico) si ricava dall’espansione delle differenze divise che può scriversi:

Risulta:

dove si è posto

**Interpolazione osculatoria di Hermite**

Siano assegnati k+1 punti reali x0,…,xk ε I a due a due distinti in corrispondenza dei quali siano noti 2k+2 valori reali f(x0),…,f(xk) e f’(x0),…,f’(xk). L’interpolazione osculatoria di Hermite consiste nel determinare un polinomio H(x) di grado al più 2k+1 tale che:

Immagine che contiene testo, orologio, calibro

Descrizione generata automaticamenteIntrodotti i polinomi di grado 2k+1

**Immagine che contiene testo

Descrizione generata automaticamente**Possiamo verificare il seguente teorema:

**Teorema: il polinomio di grado al più 2k+1**

Soddisfa le condizioni

Il polinomio H(x) è detto polinomio di interpolazione di Hermite. Il polinomio h0r(x) sono detti funzioni fondamentali di prima specie dell’interpolazione di Hermite. Il polinomio h1r(x) sono detti funzioni fondamentali di seconda specie dell’interpolazione di Hermite. Il polinomio H(x) è unico da teorema.

**Interpolazione con funzioni spline**

I due tipi di interpolazione esposti nei precedenti paragrafi presentano lo svantaggio che al crescere di k aumenta il grado del polinomio (se cresce troppo diventa ingestibile). Per ovviare a questo inconveniente si può ricorrere alla interpolazione mediante funzioni spline costituita da una interpolazione polinomiale a tratti con polinomi di uguali grado su ciascun tratto. Tutti questi polinomi devono soddisfare delle condizioni di regolarità. Siano dati k+1 punti reali x0 < x1 < … < xk in corrispondenza dei quali siano noti k+1 valori reali yi = f(xi) . Si ha la seguente definizione:

Immagine che contiene testo

Descrizione generata automaticamente**Funzione splina di grado m:** dicesi funzione splina di grado m, relativa ai punti x0,…,xk una funzione Sm(x) : [x0 , xk ] -> R tale che:

Guardiamo il caso m = 3 che dà luogo alle **spline cubiche**. Una funzione splina cubica S3(x) è composta da k polinomi pi(x), i = 1,2,…,k di grado al più 3; ciascun polinomio pi(x) : [xi-1 , xi ]-> R è definito da quattro coefficienti. S3(x) risulterà determinata dai 4k coefficienti dei polinomi che la compongono. Imponendo che siano verificate le proprietà 2 e 3, si ottengono le 4k-2 condizioni:

Immagine che contiene testo

Descrizione generata automaticamente

Immagine che contiene testo

Descrizione generata automaticamenteLe ulteriori condizioni si scelgono tra le seguenti (solitamente):

Se i valori y0’ = f’(x0) e yk’ = f’(xk) sono noti.

**Teorema**: se i punti x0,…,xk sono tali che xi – xi-1 = h, i = 1,2,…,k esiste una unica funzione spline cubica naturale S3(x).

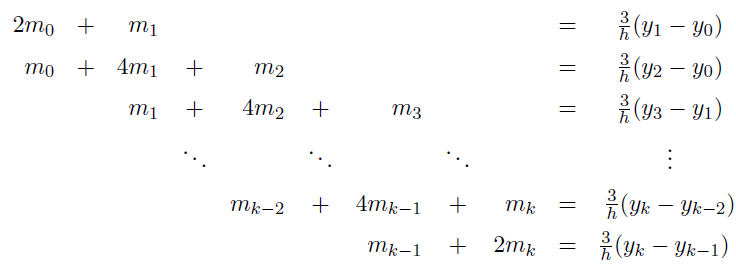
**Dimostrazione:**

Immagine che contiene testo

Descrizione generata automaticamenteintrodotti k+1 valori arbitrari mi, si costruiscono i polinomi pi(x) nella forma di Hermite a due punti con f(xr) = yr, f’(xr) = mr, r = i-1,i.

Immagine che contiene testo

Descrizione generata automaticamenteI polinomi pi(x) verificano i primi tre gruppi di condizioni per essere una funzione spline qualunque siano i valori m0,m1,…,mk. Per determinare tali k+1 valori si utilizzano le ultime condizioni delle funzioni spline e le due condizioni di spline naturale. Essendo:

si arriva al sistema lineare:

La matrice dei coefficienti di questo sistema è a predominanza diagonale forte per cui è non degenere quindi la soluzione esiste ed è unica. Esistono quindi, univocamente determinati, i polinomi p1(x),p2(x),…,pk(x).

**Immagine che contiene tavolo

Descrizione generata automaticamenteEsempio**

**Approssimazione polinomiale**

Non sempre i polinomi di interpolazione sono adatti per approssimare una funzione continua con una data accuratezza su tutto un intervallo. Infatti, per un noto teorema di Bernstein, dato un intervallo [a,b] e fissati in esso k+1 punti, esiste certamente qualche funzione f(x) continua su [a,b] con la proprietà che la successione di polinomi interpolanti P1(x), P2(x)…, di grado pari all’indice, non converge uniformemente ad f(x). Tuttavia, se per una funzione f(x) si cerca nella classe di tutti i polinomi anziché fra quelli della particolare successione sopra definita, allora esiste qualche polinomio che approssima f(x) quanto si vuole, uniformemente su [a,b]. Vale infatti il seguente teorema fondamentale.

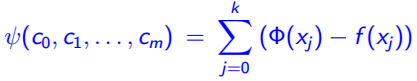
**Teorema di Weierstrass:** sia f(x) ε C0 ([a,b]); allora per ogni ε > 0 esiste un intero n, dipendente da ε, ed un polinomio p(x) di grado al più n, tale che sia

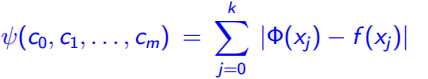
**Metodo dei minimi quadrati nel discreto**

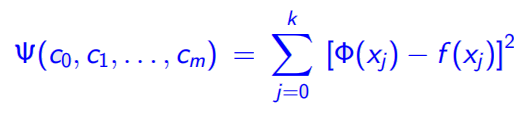
Siano dati k+1 punti xj ε I, j = 0,1,…,k e si abbiano valori f(xj) per ogni xj. Si scelgono m+1 funzioni φi(x), i = 0,1,…,m con m ≤ k, continue almeno sull’insieme I e linearmente indipendenti. Si considera la funzione combinazione lineare delle φi(x):



dove ci ε R e si vuole approssimare f(x) con φ(x). La funzione φ(x) dipende dalla scelta dei coefficienti ci. Bisogna quindi stabilire un criterio che permetta di determinare i valori ci. Poiché l’obiettivo è l’approssimazione di f(x) sembra logico mettere a confronto i valori f(xi) con i valori φ(xi) ed in particolare valutare le differenze fra i due numeri φ(xi) – f(xi).

Un’ultima possibilità potrebbe essere quella di prendere in esame la funzione

Questa però non sembra la migliore scelta perché le differenze calcolate tra i due punti sono in segno per cui si potrebbe avere una somma degli scarti uguale a zero anche avendo considerevoli differenze in alcuni punti. Quindi viene naturale pensare ad una modifica del tipo:

dove non risulta possibile l’eliminazione di alcuni addendi visto che ne è considerato solo il modulo. In questo caso l’inconveniente si presenta nella non differenziabilità della funzione in tutto lo spazio Rm+1 a cui appartiene il vettore c = (c0,c1,…,cm). Si arriva quindi a prendere in considerazione la funzione:

che assume sicuramente valori non negativi e risulta differenziabile in tutto Rm+1.

Immagine che contiene testo, orologio

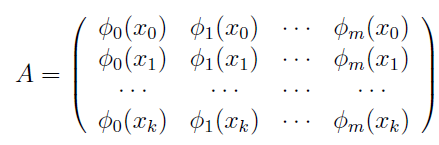
Descrizione generata automaticamenteIl metodo dei minimi quadrati consiste nello scegliere i coefficienti ci per i quali la funzione

Immagine che contiene testo, lavagnabianca

Descrizione generata automaticamenteassume valore minimo. Questa funzione rappresenta la somma degli scarti quadratici tra la funzione φ(x) e la funzione f(x) nei punti xj e coincide con il quadrato della norma euclidea del vettore dei componenti φ(xj) – f(xj). È una funzione continua di m+1 variabili, per cui il punto di minimo assoluto (se esiste) è da ricercarti tra i punti di Rm+1 per i quali sono nulle tutte le derivate parziali di ψ rispetto a c0,c1,…,cm. La funzione ψ(c ) assume valori non negativi quindi se per un c\* si avesse ψ(c\*) = 0 la funzione φ(x) risulta interpolante. Si hanno quindi le relazioni

che costituiscono un sistema lineare di m+1 equazioni nelle m+1 incognite c0,c1,…,cm.

Poniamo c= (c0,c1,..cm)T, b = ( f(x0), f(x1),…,f(xk))T,





La funzione ψ(c) può essere scritta come

Immagine che contiene testo, orologio

Descrizione generata automaticamenteIl sistema dato dalle derivate parziali uguagliato a zero, detto sistema delle equazioni normali, può scriversi nella forma: Questo sistema ha soluzione perché r(ATA) = r(ATA |ATb). Nel caso in cui la matrice A ε R(k+1) \* (m+1) abbia rango uguale a m+1, per il teorema di Binet nella sua versione estesa, la matrice ATA ε R(m+1)\*(m+1) ha determinante diverso da 0 quindi il sistema delle equazioni normali ha una sola soluzione e la funzione ψ(c) ha un unico punto stazionario (tale punto risulta essere il minimo assoluto). La scelta delle funzioni φi(x) può dipendere dalla distribuzione sul piano cartesiano dei punti di coordinate (xj, f(xj) ), j = 0,1,…,k . Le scelte più comuni sono:

Immagine che contiene testo

Descrizione generata automaticamente

**Esempio**

Calcolare l’equazione della retta e della parabola che approssimano nel senso dei minimi quadrati la funzione f(x) di cui si conoscono i seguenti valori e indicare quale delle due è da preferirsi.

**Immagine che contiene testo, parcheggio, linea

Descrizione generata automaticamenteImmagine che contiene tavolo

Descrizione generata automaticamente**

**Immagine che contiene testo

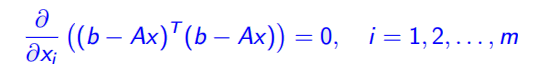
Descrizione generata automaticamente**

Risulta quindi da preferirsi l’approssimazione φ2(x)

**Sistemi lineari sovradeterminati**

Immagine che contiene testo, orologio, calibro

Descrizione generata automaticamenteSiano k e m due numeri naturali con k > m e inoltre si abbiano A ε Rk\*m e b ε Rk. Il sistema sovradeterminata Ax = b ha soluzione se e solo se r(A) = r(A|b). Nel caso in cui r(A) < r(A|b), il sistema non ha soluzione e quindi per ogni vettore x ε Rm si ha b-Ax = r dove r ε Rk, il vettore residuo risulta nullo. Si dice che si risolve il sistema nel senso dei minimi quadrati se si determina un vettore x ε Rm che renda minimo il prodotto scalare:

La determinazione di un punto di minimo è ricondotta alla ricerca dei punti che annullano tutte le derivate parziali prime della ψ(x) ossia alla risoluzione del sistema lineare

Il sistema delle derivate parziali uguagliate a zero si può scrivere nella forma: ATAx = ATb la cui risolubilità risulta da quanto si è detto per il sistema delle equazioni normali.

**Esempio**

Immagine che contiene tavolo

Descrizione generata automaticamenteSi consideri il sistema lineare sovradeterminato e si calcoli la soluzione nel senso dei minimi quadrati:

**Esempio 2**

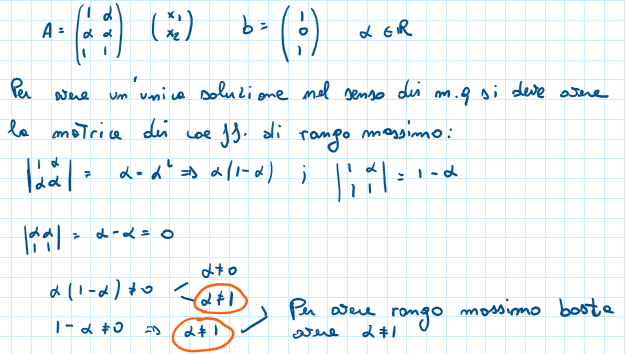
Si consideri il sistema lineare sovradeterminato. Determinare le coppie di valori di α e β per le quali il sistema non ha una soluzione unica nel senso dei minimi quadrati:

Immagine che contiene tavolo

Descrizione generata automaticamente

**Esempio**

Si consideri il sistema lineare sovradeterminato. Determinare i valori reali α per i quali il sistema ha un’unica soluzione nel senso dei minimi quadrati.

****

**Integrazione numerica**

Di molte funzioni integrabili non si conosce una funzione primitiva esprimibile con funzioni elementari. Spessa della funzione integranda è nota solo una restrizione a un insieme discreto.

**Grado di precisione ed errore**

Immagine che contiene testo

Descrizione generata automaticamenteSia f(x) sufficientemente regolare nell’intervallo [a,b] dell’asse reale. Sia ρ(x) una funzione peso non negativa in [a,b] e tale che esistano finiti i momenti:

Immagine che contiene testo, orologio

Descrizione generata automaticamente

Si pone il problema di approssimare

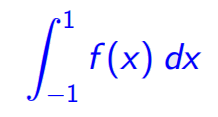
Immagine che contiene testo, orologio

Descrizione generata automaticamentecon una formula di quadratura del tipo:

dove i numeri ai (detti pesi o coefficienti) sono reali. I punti xi sono detti nodi e di solito appartengono all’intervallo [a,b]. L’errore è formalmente

Considerata la base 1,x,x2,…,xm+1 dello spazio vettoriale dei polinomi, si può dare la seguente definizione:

**Definizione**: la formula Jn(f)a ha grado di precisione (algebrico) m ε N se si verifica:

**Esempio**

Consideriamo l’integrale definito . Si vuole calcolare i pesi reali a0 e a1 della formula

per avere il massimo grado di precisione:

Immagine che contiene tavolo

Descrizione generata automaticamente

La formula trovata risulta esatta per qualunque potenza dispari di x.

**Formule di Newton-Cotes**

Immagine che contiene testo, cielo

Descrizione generata automaticamenteSe i nodi sono prefissati arbitrariamente (a due a due distinti), i pesi sono determinati dalle prime n+1 equazioni che formano il sistema lineare Vα = μ con V:

αt = (a0,a1,…,an), μT = (m0,m1,…,mn). Poiché VT è una matrice di Vandermonde, nelle ipotesi attuali che det(V) ≠ 0 e quindi α esiste ed è unico. Il grado di precisione è almeno n poiché En(1) = En(x) = … = En(xn) = 0. Se ρ(x) = 1 e i nodi sono fissati in progressione aritmetica di ragione h = (b-a)/n, cioè xi = x0 + ih, i = 0,1,…,n, e quindi con x0 = a e xn = b, si hanno le formule di Newton-Cotes; h si dice il **passo** della formula. I pesi delle formule di Newton-Cotes sono stati calcolati per vari valori di n. Fino a n = 7 (otto punti) i pesi sono positivi mentre per n > 7 i pesi sono negativi e le formule diventano numericamente instabili. Una caratteristica importante delle formule di Newton-Cotes risiede nel fatto che per esse il nucleo di Peano G(t) non cambia segno in [a,b] per cui per l’errore si può utilizzare la:

Immagine che contiene testo

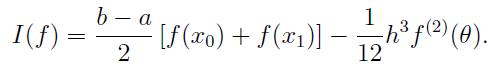
Descrizione generata automaticamente

dove m si determina in base alla definizione, mentre il termine En(xm+1) si calcola una volta noti i pesi. In tal modo, indicando con p = n+1 il numero dei nodi, l’errore assume la seguente forma caratteristica:

Immagine che contiene testo

Descrizione generata automaticamente

con cpcostante dipendente da p. Notiamo inoltre che le formule di Newton-Cotes qui definite vengono dette chiuse, per distinguerle dalle formule analoghe dette aperte nelle quali si ha: a < x0 < x1 < … < xn < b.

**Formula trapezoidale:**

**Formula di Simpson**

Se il passo di integrazione h = (b-a)/n risulta troppo ampio, si divide [a,b] in m parti uguali, mediante i punti x0 = a < x1 < x2 <… < xm = b e si utilizza la proprietà:

Immagine che contiene orologio

Descrizione generata automaticamente

Si ottengono così le formule di Newton-Cotes generalizzate. Le più usate sono la formula trapezoidale (n = Immagine che contiene testo

Descrizione generata automaticamente1) e la formula di Simpson (n = 2) le quali danno luogo alle corrispondenti formule generalizzate:

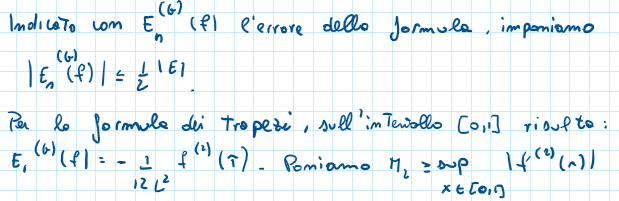
Queste formule presentano il vantaggio di un errore che tende a zero al crescere di m. Si noti che da h = (b-a)/nm segue che gli errori sono rispettivamente di ordine h2 e di h4.

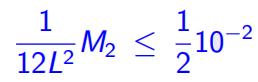
**Esempio**

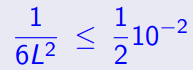
Si vuole approssimare con un massimo errore assoluto E con |E| ≤ 10-2 .

Utilizziamo la formula trapezoidale e la formula i Cavalieri-Simpson e stabiliamo quale delle due è da preferirsi. L’errore che si commette nell’approssimare un integrale con una formula di quadratura è composto da due contributi:

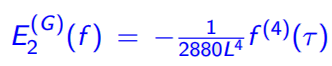
* Errore algoritmo dovuto alla sostituzione dell’integrale con un algoritmo di calcolo che è dato dalla formula considerata (coincide con l’errore della formula).
* Errore trasmesso dai dati dovuto al calcolo concreto del valore dato dalla formula. Difficile da calcolare perché dipende dagli strumenti che si hanno a disposizione.



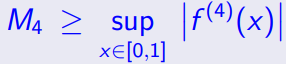
Per determinare quanti devono essere gli L intervalli per rientrare nella limitazione dell’errore si risolve la disequazione

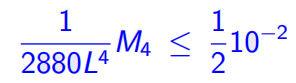


Poiché si ha: M2 = 2 quindi risolvo la disequazione le cui soluzioni sono L ≥ 6.

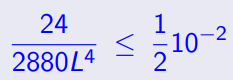
Si deduce che per rientrare nella limitazione richiesta, basta applicare la formula dei trapezi con L = 6. Il costo computazionale si misura in base a quante valutazioni di funzione si devono fare per calcolare il valore della formula di quadratura. Con L intervalli la formula dei trapezi prevede VT = L+1 valutazioni di funzioni per cui risulta VT = 7.

Per la formula di **Cavalieri-Simpson** sull’intervallo [0,1] risulta . Poniamo



Per determinare quanti devono essere gli L intervalli per rientrare nella limitazione dell’errore si risolve la disequazione **.**

Poiché



Si ha M4 = 24. Si deve quindi risolvere la disequazione le cui soluzioni sono L ≥ 2.

Quindi per rientrare nella limitazione richiesta, basta applicare la formula di Cavalieri-Simpson con L = 2. Il costo computazione si misura in base a quante valutazioni di funzioni si devono fare per calcolare il valore della formula di quadratura. Con L intervalli, la formula di Cavalieri-Simpson prevede VCS = 2L +1 quindi nel nostro caso risulta VCS = 5

**Formule di quadratura di tipo Gaussiano**

Sia ρ(x) una funzione peso che verifichi le ipotesi:

Immagine che contiene testo, orologio, calibro

Descrizione generata automaticamenteSi indichi con Π lo spazio vettoriale dei polinomi algebrici a coefficienti reali e sia [a,b] un intervallo, non necessariamente limitato. Per ogni coppia r(x),s(x) ε Π su consideri il prodotto scalare:

Si definisce quindi la classe dei polinomi ortogonali Π\* come l’insieme dei polinomi algebrici, a coefficienti reali, ortogonali rispetto al prodotto scalare, cioè:

Immagine che contiene testo

Descrizione generata automaticamenteI numeri positivi hi sono le costanti di normalizzazione. Dati [a,b] e ρ(x), gli elementi di Π\* restano definiti a meno di una costante moltiplicativa e costituiscono una base per Π; per essi valgono le seguenti proprietà:

Immagine che contiene testo, orologio

Descrizione generata automaticamenteFissati ρ(x), [a,b] e n risulta univocamente determinata la formula di quadratura di grado di precisione almeno 2n+1

Chiamata formula di quadratura gaussiana.