Calcolo Numerico

Metodi per equazioni e sistemi non lineari

Simone Rebegoldi

Corso di Laurea in Informatica
Dipartimento di Scienze Fisiche, Informatiche e Matematiche

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Optimization Algorithms and Software for Inverse problemS

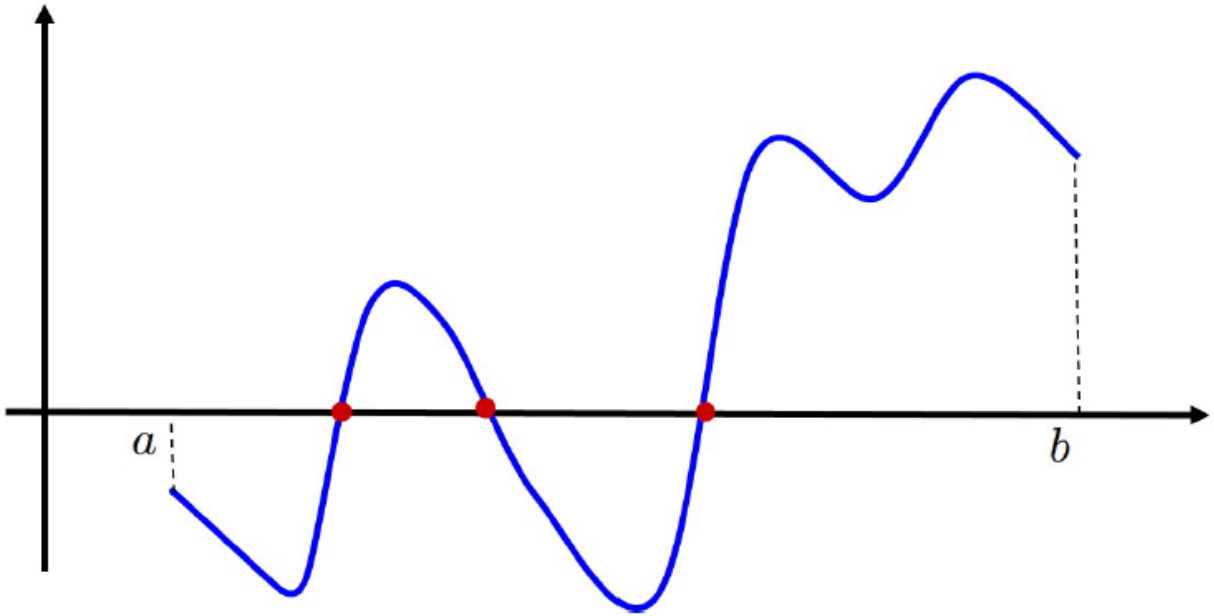
www.oasis.unimore.it

1. Metodi per equazioni non lineari

Definizione

Sia $f : [a, b] \to \mathbb{R}$ una funzione. Il punto $x_* \in [a, b]$ si dice radice (o zero) della funzione f se

$$f(x_*) = 0.$$



Teorema di esistenza degli zeri

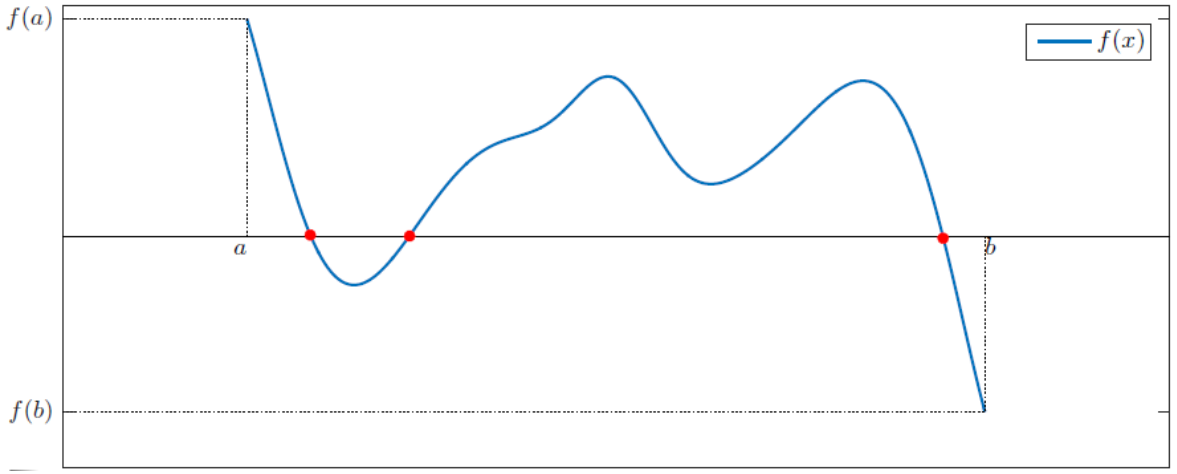
Sia $f : [a, b] \to \mathbb{R}$ una funzione continua tale che

$$f(a)f(b) < 0.$$

Allora esiste almeno uno zero di f nell'intervallo $[a, b]$, ossia esiste $x_* \in [a, b]$ tale che

$$f(x_*) = 0.$$

- Le ipotesi del teorema precedente non garantiscono l'unicità della soluzione. Una condizione sufficiente per avere l'unicità è, ad esempio, la stretta monotonia della funzione



Problema del condizionamento

Sia x_* una radice di $f : [a, b] \to \mathbb{R}$, ossia $f(x_*) = 0$.
Sia $\tilde{x}$ una soluzione del problema perturbato $f(x) = \delta$ con δ "piccolo".
Sotto quali condizioni possiamo concludere che $|x_* - \tilde{x}|$ è altrettanto "piccolo"?

- Per semplicità, assumiamo che la funzione f di cui vogliamo calcolare le radici sia derivabile. Dalla definizione di derivata, si ha

$$f'(x_*) = \lim_{x \to x_*} \frac{f(x) - f(x_*)}{x - x_*}.$$

  Dunque, in un intorno di x_*, si può operare la seguente approssimazione

$$f'(x_*) \simeq \frac{f(x) - f(x_*)}{x - x_*}. \tag{1}$$

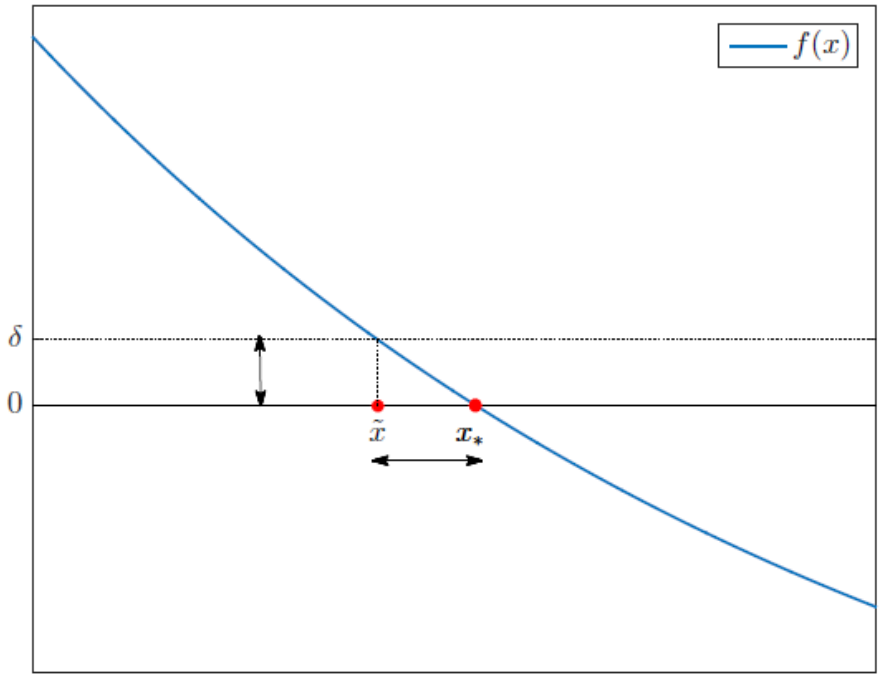
- Valutando l'approssimazione (1) in $x = \tilde{x}$ e passando ai valori assoluti, si ha

$$|x_* - \tilde{x}| \simeq \frac{|f(x_*) - f(\tilde{x})|}{|f'(x_*)|} = \frac{\delta}{|f'(x_*)|}.$$

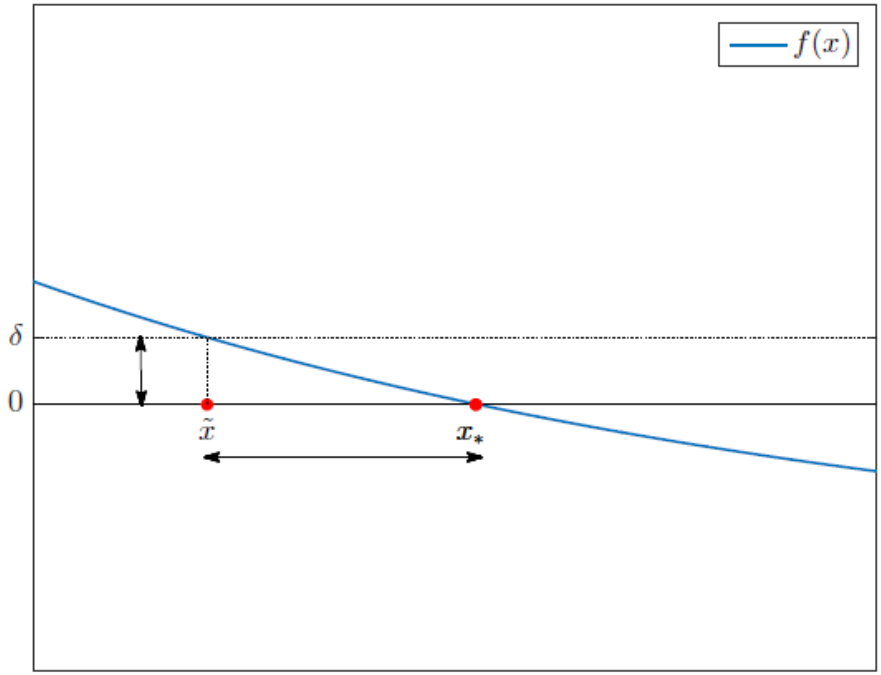
  $\Rightarrow$ Se $f'(x_*) \simeq 0$, allora la soluzione $\tilde{x}$ del problema perturbato potrebbe essere distante dalla radice x_*, ossia $|\tilde{x} - x_*|$ potrebbe essere grande, anche se δ è piccolo.

- Il problema $f(x) = 0$ è mal condizionato quando $f'(x_*) \simeq 0$.
- La condizione $f'(x_*) \simeq 0$ implica che il grafico di f risulti "appiattito" sull'asse orizzontale in un intorno di x_*.

Problema ben condizionato



Problema mal condizionato



- Se f è non lineare, ovvero $f(x)$ non è esprimibile come combinazione lineare delle componenti di x, allora gli zeri di f non sono generalmente esprimibili in forma chiusa.

  Esempio: $f(x) = xe^x$ ammette almeno una soluzione in $[-1, 1]$, ma non esiste una formula analitica per calcolarla.

- Di conseguenza, è necessario fare ricorso ai metodi iterativi.

- Studieremo i seguenti metodi:
  1. Metodo di bisezione
  2. Metodo di Newton
  3. Metodo delle secanti
  4. Metodo del punto fisso (o delle approssimazioni successive).

Dati

Intervallo di ricerca $[a, b]$.

Ipotesi

Funzione continua che assume segno discorde agli estremi dell'intervallo: $f(a)f(b) < 0$.

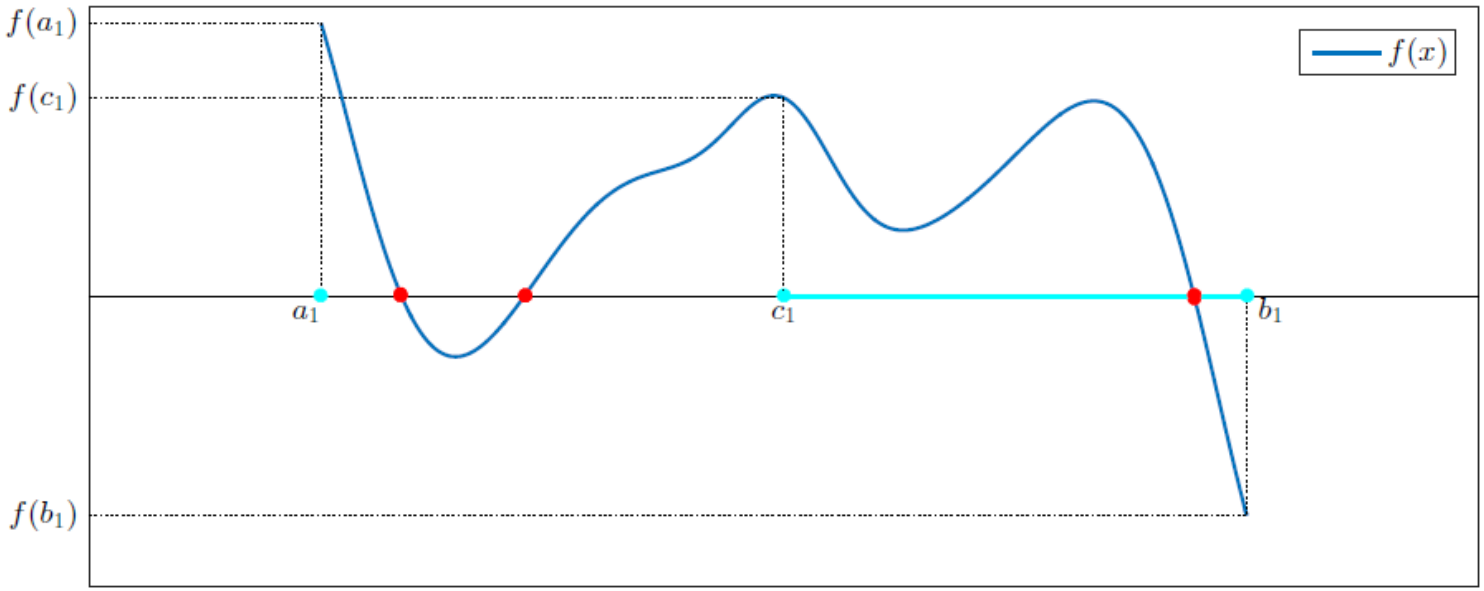
Descrizione

Si applica ripetutamente il teorema di esistenza degli zeri, generando una successione di intervalli di ampiezza decrescente i cui punti medi convergono ad uno zero di f.

Passo 1

- Si calcola il punto medio c_1 dell'intervallo di ricerca $[a_1, b_1] \equiv [a, b]$.
- Si definisce il nuovo intervallo di ricerca $[a_2, b_2]$ come quello tra i due sottointervalli $[a_1, c_1]$, $[c_1, b_1]$ in cui sono soddisfatte le ipotesi del teorema di esistenza degli zeri:

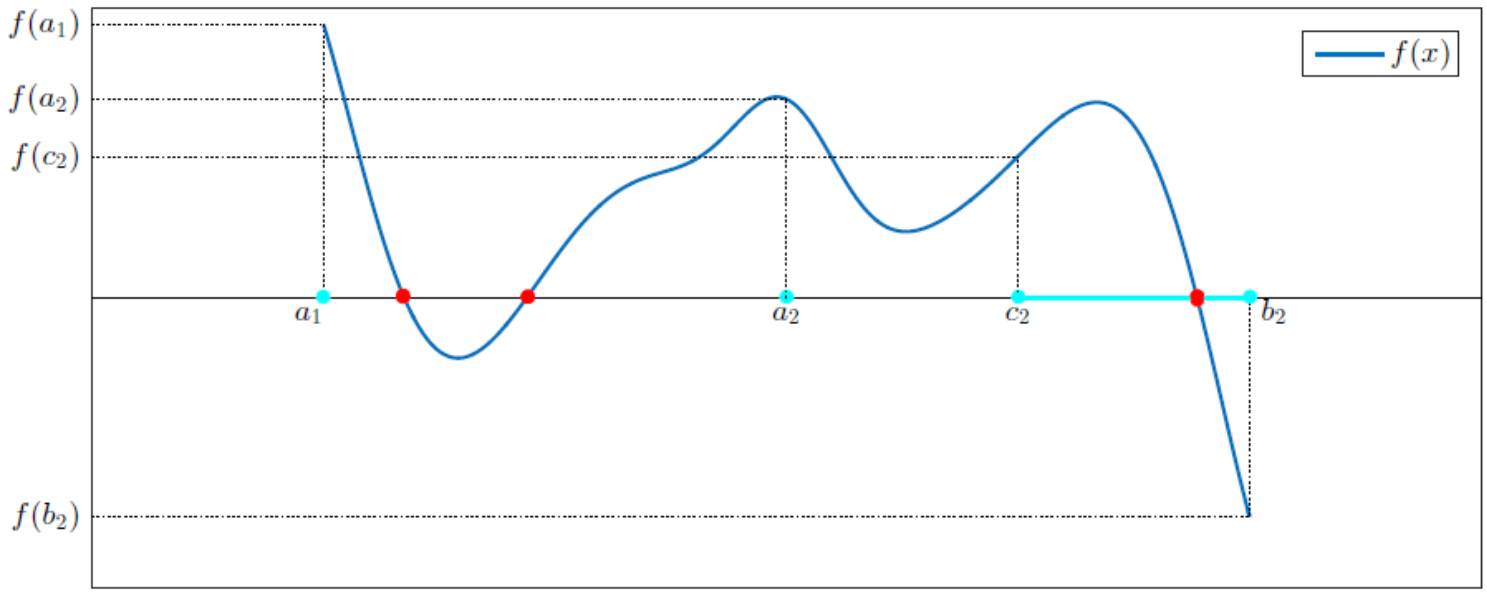
$$[a_2, b_2] = \begin{cases} [a_1, c_1], & \text{se } f(c_1)f(a_1) < 0 \\ [c_1, b_1], & \text{se } f(c_1)f(b_1) < 0. \end{cases}$$



Passo 2

- Si calcola il punto medio c_2 dell'intervallo di ricerca $[a_2, b_2]$.
- Si definisce il nuovo intervallo di ricerca $[a_3, b_3]$ come quello tra i due sottointervalli $[a_2, c_2]$, $[c_2, b_2]$ in cui sono soddisfatte le ipotesi del teorema di esistenza degli zeri:

$$[a_3, b_3] = \begin{cases} [a_2, c_2], & \text{se } f(c_2)f(a_2) < 0 \\ [c_2, b_2], & \text{se } f(c_2)f(b_2) < 0. \end{cases}$$



Passo k

- Si calcola il punto medio c_k dell'intervallo di ricerca $[a_k, b_k]$.
- Si definisce il nuovo intervallo di ricerca $[a_{k+1}, b_{k+1}]$ come quello tra i due sottointervalli $[a_k, c_k]$, $[c_k, b_k]$ in cui sono soddisfatte le ipotesi del teorema di esistenza degli zeri:

$$[a_{k+1}, b_{k+1}] = \begin{cases} [a_k, c_k], & \text{se } f(c_k)f(a_k) < 0 \\ [c_k, b_k], & \text{se } f(c_k)f(b_k) < 0. \end{cases}$$

Proprietà degli intervalli di ricerca

1. Per ogni k, le ipotesi del teorema di esistenza degli zeri sono verificate, dunque esiste almeno una radice di f in $[a_k, b_k]$.
2. Per ogni k, l'ampiezza del $k-$esimo intervallo di ricerca è data da

$$b_k - a_k = \frac{b-a}{2^{k-1}}.$$

Teorema (convergenza del metodo di bisezione)

Sia $x_* \in [a, b]$ uno zero di f.
La successione dei punti medi $\{c_k\}_{k\in\mathbb{N}}$ generata dal metodo di bisezione converge ad uno zero di f nell'intervallo $[a, b]$ e soddisfa la disuguaglianza

$$|c_k - x_*| \leq \frac{b-a}{2^{k-1}}, \quad k = 0, 1, \ldots$$

Dimostrazione

Per costruzione $f(a_k)f(b_k) < 0$ per ogni k, quindi $x_* \in [a_k, b_k]$ per ogni k.
Di conseguenza c_k e x_* stanno nello stesso intervallo, il che implica che

$$|c_k - x_*| \leq b_k - a_k = \frac{b-a}{2^{k-1}} \underset{k\to\infty}{\longrightarrow} 0.$$

Per il teorema dei carabinieri, concludiamo che

$$\lim_{k\to\infty} |c_k - x_*| = 0,$$

dunque il metodo converge. □

Osservazioni sulla convergenza

- La convergenza del metodo di bisezione è globale, nel senso che il metodo converge qualunque sia la scelta dell'intervallo $[a, b]$ tale che $f(a)f(b) < 0$.
- Fissata una tolleranza $\tau > 0$, è possibile ricavare il numero di passi sufficiente per ottenere un'approssimazione di uno zero di f entro la tolleranza τ. Infatti

$$\frac{b-a}{2^{k-1}} \leq \tau \quad \Leftrightarrow \quad k \geq 1 + \log_2\left(\frac{b-a}{\tau}\right).$$

Pertanto, usando il teorema precedente, per ogni $k \geq \lceil 1 + \log_2\left(\frac{b-a}{\tau}\right) \rceil$ si ha

$$|c_k - x_*| \leq \frac{b-a}{2^{k-1}} \leq \tau,$$

ossia c_k approssima uno zero di f entro la tolleranza τ.

Formula stabile per il calcolo del punto medio

Per calcolare il punto medio di $[a, b]$, esistono due possibili algoritmi:

$$\frac{a+b}{2}, \quad a + \frac{b-a}{2}.$$

Operando in aritmetica finita, il primo algoritmo è meno stabile del secondo quando a e b sono vicini tra loro.

Esempio

Dati: $a = 0.983$, $b = 0.986$, aritmetica decimale con 3 cifre di precisione e troncamento.

- Primo algoritmo

$$fl(a+b) = fl(1.969) = 0.196\ 10^{1}$$

$$fl\left(\frac{fl(a+b)}{2}\right) = 0.98. \quad \text{Punto esterno all'intervallo!}$$

- Secondo algoritmo

$$fl(b-a) = 0.3\ 10^{-2}$$

$$fl\left(\frac{b-a}{2}\right) = 0.15\ 10^{-2}$$

$$fl\left(a + fl\left(\frac{b-a}{2}\right)\right) = fl(0.983 + 0.0015) = fl(0.9845) = 0.984.$$

Algoritmo basato sul metodo di bisezione

INPUT: a, b, f, τ
$N \leftarrow \lceil 1 + \log_2((b-a)/\tau) \rceil$
$fa \leftarrow f(a)$
$fb \leftarrow f(b)$
FOR $k = 1, 2, ..., N$
  $c \leftarrow a + (b-a)/2$
  $fc \leftarrow f(c)$
  IF $f(c) = 0$
    return c
  END
  IF $f(c)f(b) < 0$
    $a \leftarrow c$
    $fa \leftarrow fc$
  ELSE
    $b \leftarrow c$
    $fb \leftarrow fc$
  END

Complessità computazionale

- Si misura in numero di valutazioni di funzione per iterazione, ossia quante volte nella singola iterazione di un metodo viene calcolata la funzione f.
- Si assume infatti che il calcolo (approssimato!) di una qualsiasi funzione non lineare (trigonometrica, esponenziale,...) si ottenga con algoritmo numerico basato su una successione di operazioni aritmetiche fondamentali.
- Ogni valutazione di funzione equivale ad un "pacchetto" di operazioni fondamentali. Il costo computazionale di una iterazione si ottiene contando i "pacchetti" invece delle singole operazioni.
- Dunque il costo del metodo di bisezione è pari ad una valutazione di funzione per iterazione.

- Si può assumere che, per ogni k, valga l'approssimazione

$$|c_k - x_*| \simeq \frac{b-a}{2^{k-1}}.$$

  Di conseguenza

$$|c_{k+1} - x_*| \simeq \frac{1}{2}|c_k - x_*|.$$

  Dunque l'errore commesso si dimezza ad ogni iterazione.
- Ciò significa che, nel metodo di bisezione, si guadagna meno di una cifra decimale di precisione ogni 3 iterazioni.
- Quantifichiamo questo fatto con il concetto di ordine di convergenza.

Definizione

Sia $\{x_k\}_{k\in\mathbb{N}} \subseteq \mathbb{R}$ una successione che converge ad un punto $x_* \in \mathbb{R}$. Si dice che la successione $\{x_k\}_{k\in\mathbb{N}}$ ha ordine di convergenza p se

$$\lim_{k\to\infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|^p} = C,$$

per qualche $p \geq 1$, $C \in \mathbb{R}$.
Se la successione è generata da un metodo iterativo, si dice che il metodo ha ordine p.

- L'ordine di convergenza permette di valutare il guadagno che si ottiene in termini di riduzione dell'errore ad ogni iterazione di un metodo iterativo.
- Se un metodo iterativo ha ordine p, applicando la definizione di limite, per k grande si ha

  $$|x_{k+1} - x_*| \simeq C|x_k - x_*|^p.$$

  Dunque più p è grande, maggiore sarà la riduzione dell'errore da una iterazione all'altra.

Casi particolari

- Convergenza quadratica ($p = 2$)

$$\lim_{k\to\infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|^2} = C.$$

- Convergenza lineare ($p = 1$, $C \in (0, 1)$)

$$\lim_{k\to\infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} = C.$$

- Convergenza superlineare ($p = 1$, $C = 0$)

$$\lim_{k\to\infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} = 0.$$

Ordine di convergenza del metodo di bisezione

Assumendo che valga l'approssimazione $|c_k - x_*| \simeq \frac{b-a}{2^k}$, si può mostrare che il metodo di bisezione ha convergenza lineare, infatti:

$$\lim_{k\to\infty} \frac{|c_{k+1} - x_*|}{|c_k - x_*|} \simeq \lim_{k\to\infty} \frac{\frac{b-a}{2^k}}{\frac{b-a}{2^{k-1}}} = \frac{1}{2}.$$

Vantaggi del metodo di bisezione

- Converge globalmente (qualunque sia la scelta dell'intervallo iniziale).
- Richiede come ipotesi soltanto la continuità della funzione.
- Ha una bassa complessità computazionale (1 valutazione di funzione per iterazione).

Svantaggi del metodo di bisezione

- Converge lentamente (linearmente) ad una soluzione
- Non si può estendere al caso di sistemi di equazioni non lineari.

Metodo di Newton (o metodo delle tangenti)

Dati

Punto iniziale $x_0 \in [a, b]$.

Ipotesi

Funzione f derivabile in $[a, b]$.

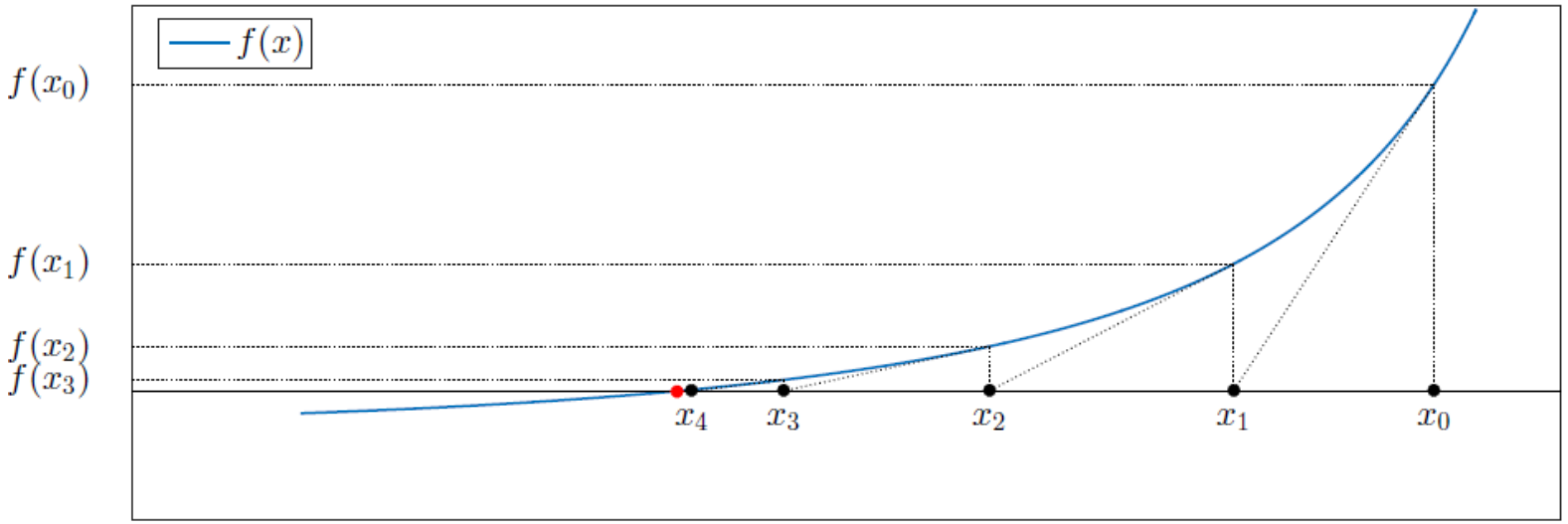
Descrizione

Al passo k, si considera la retta tangente al grafico di f nel punto $(x_k, f(x_k))$ e se ne calcola il punto di intersezione con l'asse delle ascisse, ottenendo così x_{k+1}.

Passo k

A partire dall'iterata corrente x_k, l'iterata successiva x_{k+1} viene calcolata come l'intersezione tra l'asse delle ascisse e la retta tangente al grafico di f nel punto $(x_k, f(x_k))$.

$$\begin{cases} y = 0 \\ y = f(x_k) + f'(x_k)(x - x_k) \end{cases} \quad \Rightarrow \quad x = x_k - \frac{f(x_k)}{f'(x_k)}.$$



Definizione del metodo

Dato il punto iniziale $x_0 \in [a, b]$, il metodo di Newton genera una successione di iterate $\{x_k\}_{k \in \mathbb{N}}$ della forma

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad \text{dove } f'(x_k) \neq 0.$$

Osservazioni sul metodo e sua complessità

- Affinché il metodo sia ben posto, deve essere $f'(x_k) \neq 0$ per ogni k.
  Dal punto di vista geometrico, ciò significa che non è possibile eseguire il passo di Newton se l'iterata corrente è un punto a tangente orizzontale.
- Sia data la formula di Taylor del prim'ordine con resto di Peano centrata in x_k:

$$f(x) = \underbrace{f(x_k) + f'(x_k)(x - x_k)}_{r(x) = \text{ retta tangente}} + o(|x - x_k|), \quad \forall\, x \in \mathbb{R}.$$

  Tale formula implica che la retta tangente al grafico di f in $(x_k, f(x_k))$ è una "buona" approssimazione di f in un intorno I di x_k, ossia

$$f(x) \simeq f(x_k) + f'(x_k)(x - x_k), \quad \text{per } x, x_k \in I.$$

  Quindi il passo k del metodo di Newton può essere interpretato come segue: l'equazione non lineare $f(x) = 0$ (difficile) viene sostituita con l'equazione lineare $f(x_k) + f'(x_k)(x - x_k) = 0$ (facile), ottenuta approssimando f con il suo sviluppo di Taylor centrato in x_k e troncato al primo ordine.
- Il costo computazionale per iterazione è di 2 valutazioni di funzione ($f(x_k)$ e $f'(x_k)$). Quindi il metodo di Newton è più costoso del metodo di bisezione.

Osservazione

Il metodo di Newton può oscillare senza convergere ad uno zero di f se il punto x_0 non è "sufficientemente vicino" alla soluzione del problema.

Controesempio

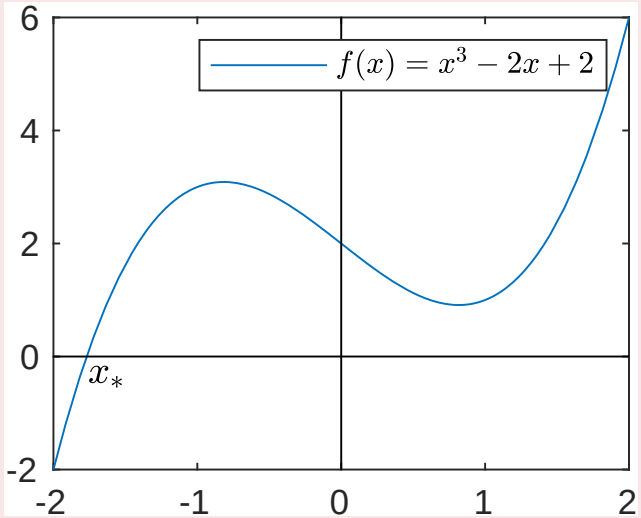
Sia $f(x) = x^3 - 2x + 2$ il cui unico zero è $x_* = -1.7693....$ e $f'(x) = 3x - 2$.
Se prendiamo $x_0 = 0$, risulta che

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 0 - \frac{2}{-2} = 1$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 1 - \frac{1}{1} = \underbrace{0}_{x_0}$$

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} = x_0 - \frac{f(x_0)}{f'(x_0)} = 1$$

$$\vdots$$



Dunque la successione $\{x_k\}_{k\in\mathbb{N}}$ oscilla tra 0 e 1 senza mai convergere a x_*.

Teorema di convergenza del metodo di Newton

Supponiamo che $f \in C^2([a,b])$, ossia f derivabile due volte con continuità.
Sia $x_* \in (a,b)$ con $f(x_*) = 0$ e $f''(x_*) \neq 0$.
Allora si può provare che:

1. esiste $\delta > 0$ tale che se $|x_0 - x_*| < \delta$, la successione $\{x_k\}_{k\in\mathbb{N}}$ generata dal metodo di Newton converge a x_*;
2. se il metodo converge a x_*, allora

$$\lim_{k\to\infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|^2} = C, \quad C \in \mathbb{R}.$$

- Il punto 1 ci dice che il metodo di Newton ha convergenza locale, ossia converge soltanto se x_0 è "sufficientemente vicino" ad uno zero di f.
- Il punto 2 ci dice che il metodo di Newton, se converge, ha convergenza quadratica (di un ordine superiore a quella di bisezione).

Dimostrazione (punto 2)

Applichiamo la formula di Taylor del secondo ordine con resto di Lagrange:

$$f(x_*) = f(x_k) + f'(x_k)(x_* - x_k) + \frac{1}{2}f''(\xi_k)(x_* - x_k)^2, \quad \xi_k \in [x_k, x_*].$$

Siccome $f(x_*) = 0$, dividendo entrambi i membri della precedente uguaglianza per $f'(x_k)$ si ottiene

$$0 = \underbrace{\frac{f(x_k)}{f'(x_k)} - x_k}_{=-x_{k+1}} + x_* + \frac{f''(\xi_k)}{2f'(x_k)}(x_* - x_k)^2$$

$$0 = x_* - x_{k+1} + \frac{f''(\xi_k)}{2f'(x_k)}(x_* - x_k)^2.$$

Dividendo entrambi i membri per $(x_* - x_k)^2$ si ottiene

$$\frac{x_* - x_{k+1}}{(x_* - x_k)^2} = -\frac{f''(\xi_k)}{2f'(x_k)}.$$

Infine, passando al limite e applicando la continuità di f' e f'', si ottiene

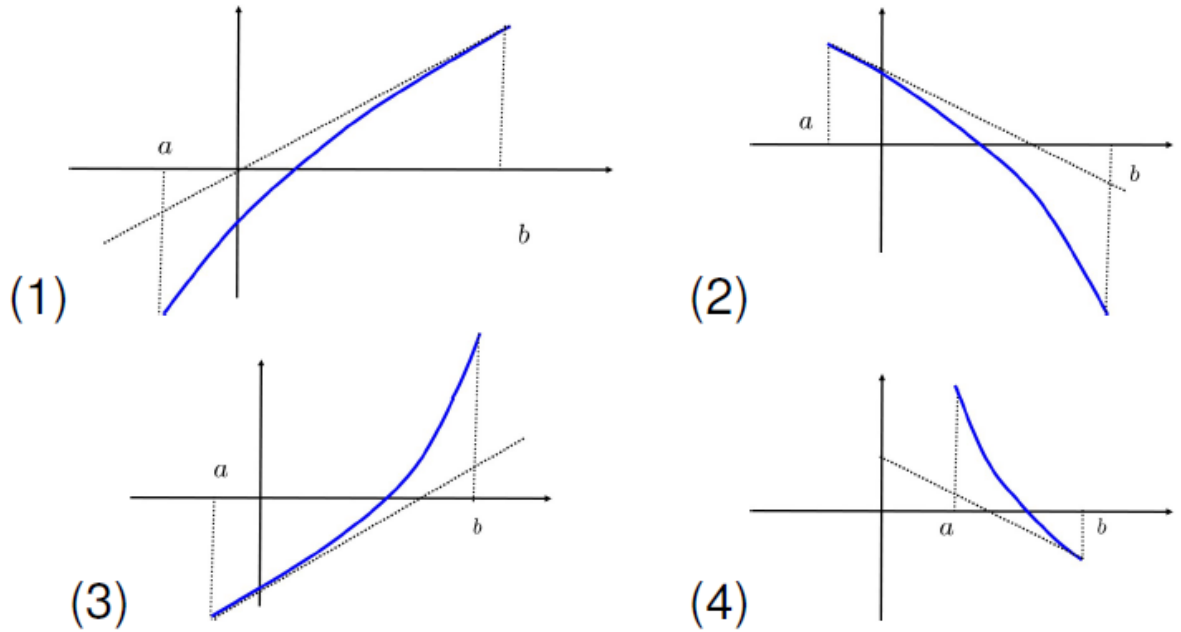
$$\lim_{k\to\infty} \frac{|x_* - x_{k+1}|}{|x_* - x_k|^2} = \frac{|f''(x_*)|}{|2f'(x_*)|}. \quad \square$$

Condizioni sufficienti per la convergenza del metodo di Newton

Sia $f \in C^2([a,b])$ tale che $f(a)f(b) < 0$. Siano soddisfatte le seguenti ipotesi:

- il segno di $f'(x)$ è costante su $[a,b]$ ($f'(x) > 0$ o $f'(x) < 0$ per ogni $x \in [a,b]$);
- il segno di $f''(x)$ è costante su $[a,b]$ ($f''(x) > 0$ o $f''(x) < 0$ per ogni $x \in [a,b]$).

Allora se $x_0 \in [a,b]$ è tale che $f(x_0)f''(x_0) > 0$, la successione $\{x_k\}_{k\in\mathbb{N}}$ generata dal metodo di Newton converge all'unico zero di f in $[a,b]$.

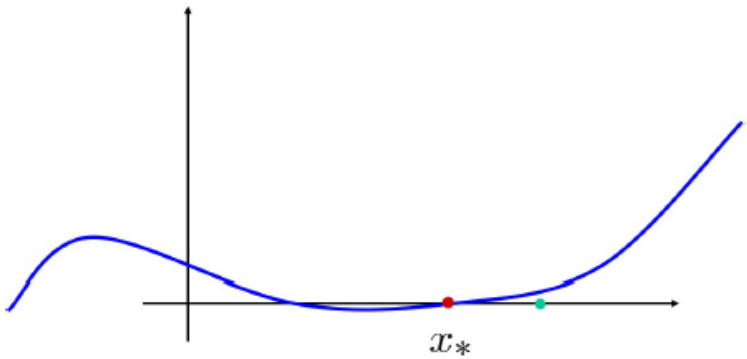


Criteri di arresto

- Ad eccezione del metodo di bisezione, nei metodi iterativi non si conosce a priori il numero di iterazioni sufficiente ad ottenere un'approssimazione della soluzione entro una tolleranza fissata.
- Solitamente si adotta una combinazione di criteri di arresto, analoghi a quelli usati per i metodi iterativi per sistemi lineari, basati su due quantità:
  - la differenza tra due iterate successive;
  - il residuo del problema, definito come una quantità che si annulla in corrispondenza della soluzione.
- Fissata una tolleranza $\tau > 0$, i metodi per la ricerca degli zeri di funzione si arrestano alla prima iterazione k in cui sono soddisfatte le disuguaglianze

$$\frac{|x_{k+1} - x_k|}{|x_{k+1}|} \leq \tau \quad \text{e} \quad |f(x_k)| \leq \tau.$$

  Notare che la condizione $f(x_k) \leq \tau$ è poco affidabile quando il grafico di f è molto schiacciato sull'asse x.



Algoritmo basato sul metodo di Newton

INPUT: $f, f', x, \tau, N_{\max}$
FOR $k = 1, 2, ..., N_{\max}$
  $fx \leftarrow f(x)$
  $dfx \leftarrow f'(x)$
  IF $dfx = 0$
    print warning and return
  END
  $x_{new} \leftarrow x - fx/dfx$
  IF $|x - x_{new}|/|x_{new}| \leq \tau$ AND $|fx| \leq \tau$
    set $x = x_{new}$ and return x
  END
  $x \leftarrow x_{new}$
END
OUTPUT: x

Vantaggi del metodo di Newton

- Se converge, ha convergenza quadratica (più veloce della bisezione).
- Si può estendere al caso di sistemi di equazioni non lineari.

Svantaggi del metodo di Newton

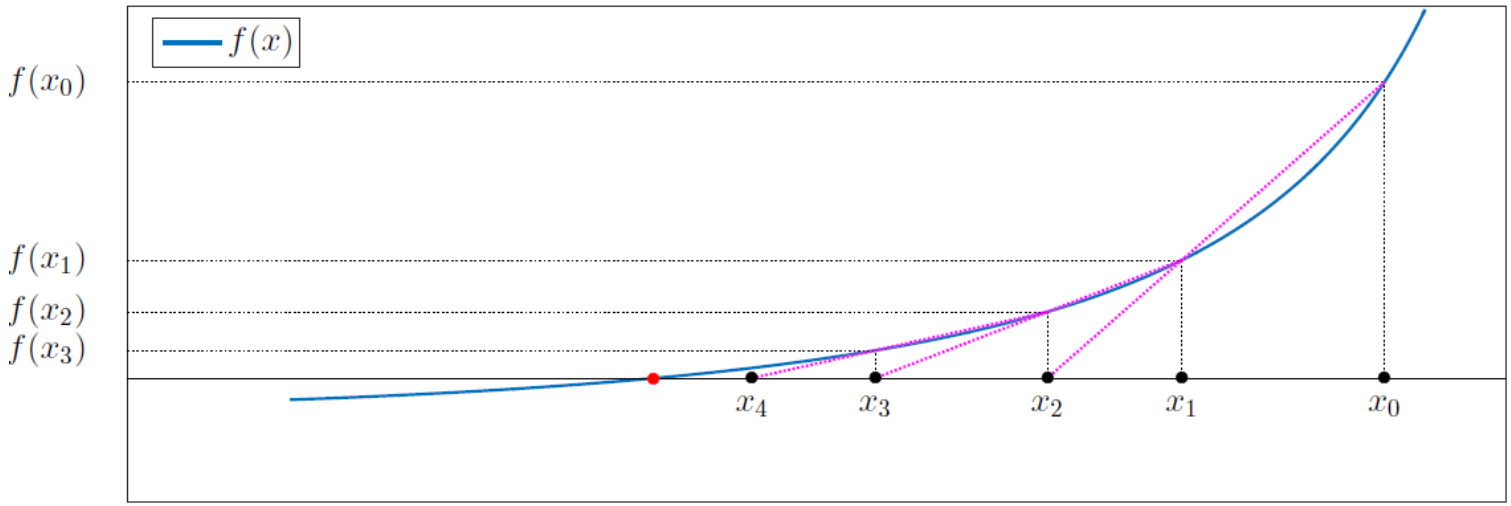
- Converge soltanto localmente (se x_0 è abbastanza vicino ad una soluzione).
- Ha una maggiore complessità computazionale rispetto alla bisezione (2 valutazioni di funzione per iterazione).

È una variante del metodo di Newton in cui la derivata prima viene approssimata con un rapporto incrementale:

$$x_{k+1} = x_k - \frac{f(x_k)}{\frac{f(x_k)-f(x_{k-1})}{x_k - x_{k-1}}}, \quad k = 0, 1, \ldots$$

L'iterata successiva x_{k+1} è l'intersezione tra l'asse delle ascisse e la retta passante per i punti $(x_{k-1}, f(x_{k-1}))$ e $(x_k, f(x_k))$.

- Non richiede il calcolo della derivata prima.
- Se ne può dimostrare, sotto opportune ipotesi, la convergenza superlineare.
- Ne esiste un'estensione per sistemi non lineari, detto metodo Quasi-Newton.



Il metodo del punto fisso (o metodo delle approssimazioni successive) può essere ricavato operando un'analogia con i metodi iterativi per i sistemi lineari.

Sistema lineare	Equazione non lineare
$Ax = b \Leftrightarrow b - Ax = 0$	$f(x) = 0$
$0 = -M^{-1}(Ax - b)$ dove M è non singolare	$-\phi(x)f(x) = 0$ dove $\phi : \mathbb{R} \to \mathbb{R},\ \phi(x) \neq 0\ \forall\ x$
$x = x - M^{-1}(Ax - b)$	$x = x - \phi(x)f(x)$
$x = \underbrace{(I - M^{-1}A)}_{G}x + \underbrace{M^{-1}b}_{c}$	$x = \underbrace{x - \phi(x)f(x)}_{g(x)}$
$x = Gx + c$	$x = g(x)$

Metodo iterativo per $Ax = b$

$$x^{(k+1)} = Gx^{(k)} + c$$

Metodo del punto fisso per $f(x) = 0$

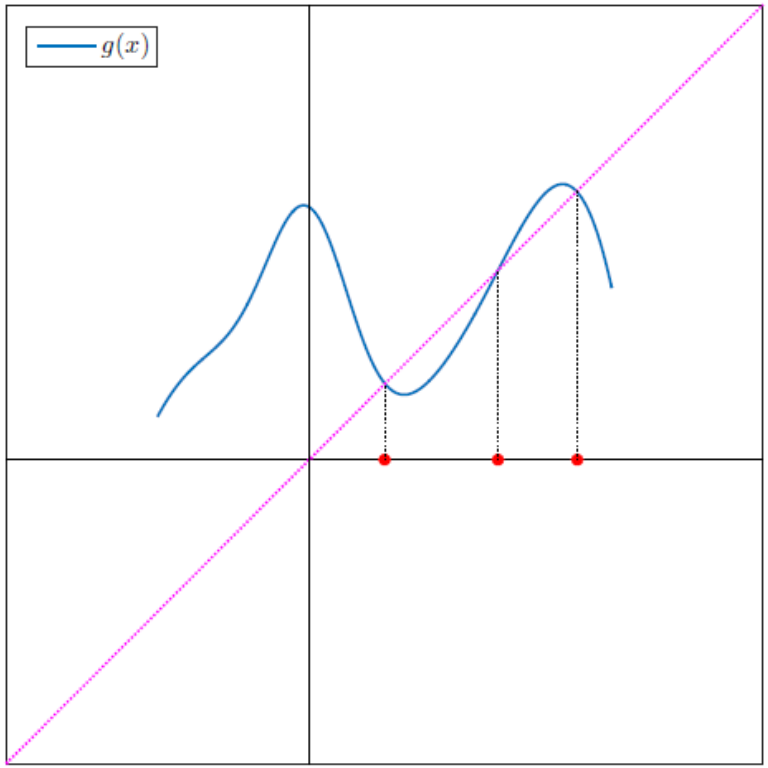
$$x_{k+1} = g(x_k)$$

Definizione

Data una funzione $g : [a, b] \to \mathbb{R}$, un punto $x_* \in [a, b]$ è detto punto fisso della funzione g se

$$g(x_*) = x_*.$$

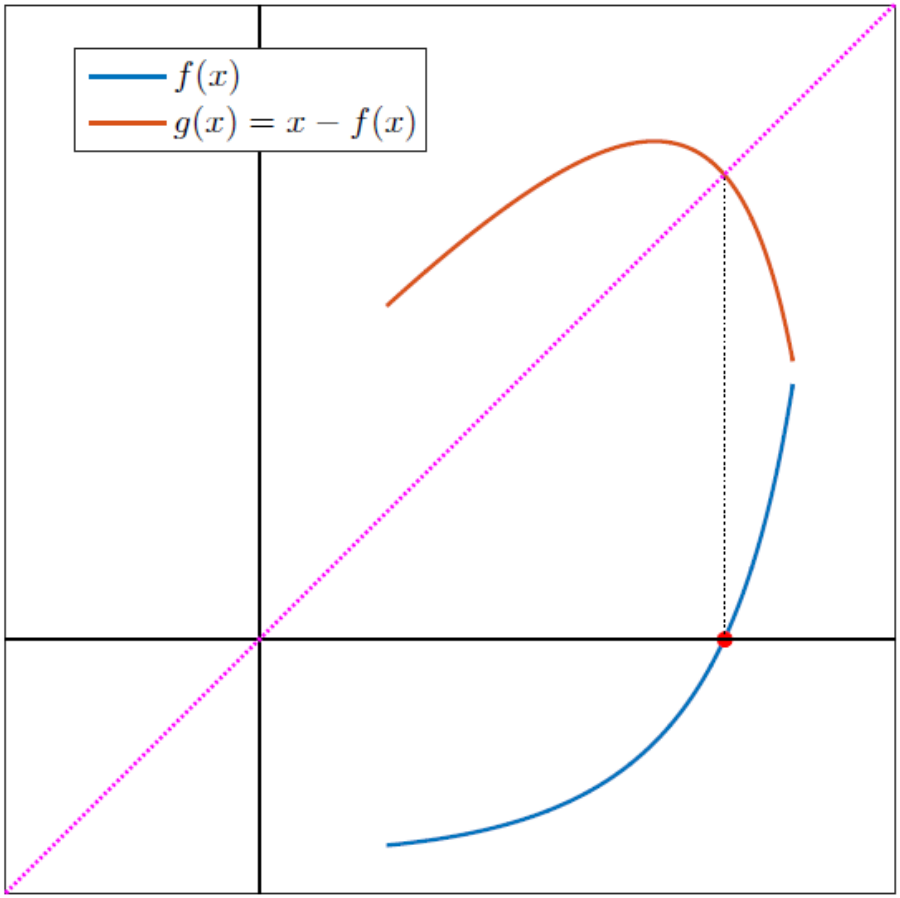
- Dal punto di vista geometrico, un punto fisso di g corrisponde ad un punto in cui il grafico della funzione g interseca la bisettrice del primo e del terzo quadrante, avente equazione $y = x$.



Proposizione

Sia $f : [a, b] \to \mathbb{R}$. Data una funzione $\phi : [a, b] \to \mathbb{R}$ con $\phi(x) \neq 0 \; \forall \; x \in [a, b]$, si ha che gli zeri di f sono tutti e soli i punti fissi della funzione

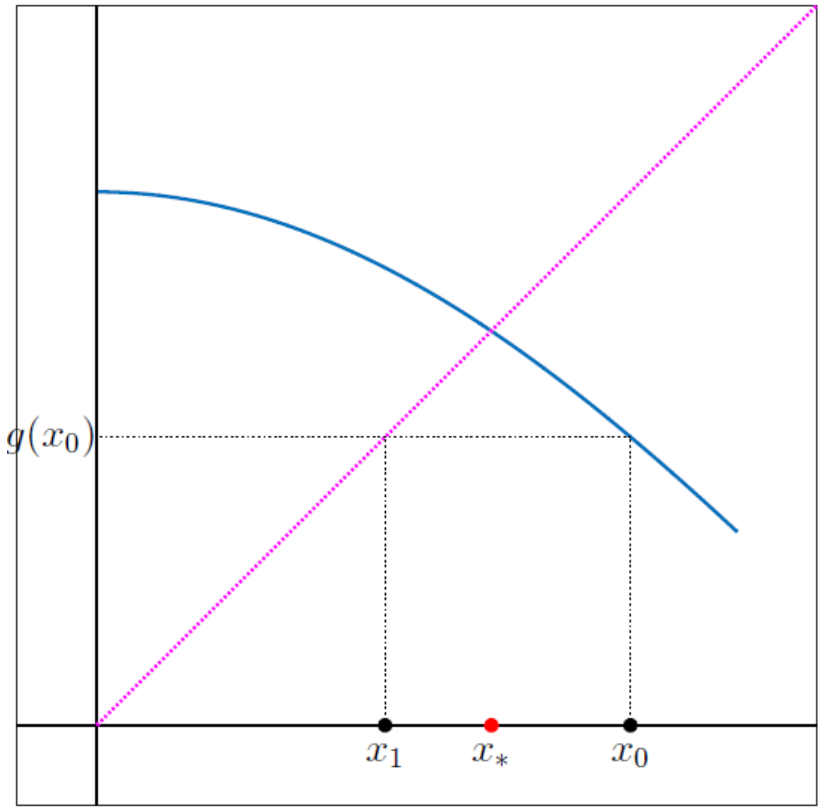
$$g(x) = x - \phi(x) f(x).$$



Definizione

Data una funzione $g : [a, b] \to \mathbb{R}$ e dato un punto iniziale $x_0 \in [a, b]$, il metodo del punto fisso, detto anche metodo delle approssimazioni successive, è definito dalla seguente iterazione

$$x_{k+1} = g(x_k), \quad k = 0, 1, \ldots$$



- Come nel caso dei sistemi lineari, in cui la scelta di M corrisponde ad un diverso metodo, la scelta della funzione $\phi(x)$ nella definizione di $g(x) = x - \phi(x)f(x)$ determina diversi metodi delle approssimazioni successive che sono anche metodi per la ricerca degli zeri di f.
- Sotto opportune ipotesi, scegliendo

$$\phi(x) = \frac{1}{f'(x)}$$

il metodo delle approssimazioni successive per la ricerca dei punti fissi di $g(x) = x - \phi(x)f(x)$ corrisponde al metodo di Newton applicato alla ricerca degli zeri di f.

Teorema della mappa contrattiva

Sia $g : [a.b] \to \mathbb{R}$ con $g(x) \in [a, b] \; \forall \; x \in [a, b]$.
Supponiamo che g sia una contrazione in $[a, b]$, ossia esiste $L \in (0, 1)$ tale che

$$|g(x) - g(y)| \leq L|x - y|, \quad \forall \; x, y \in [a, b].$$

Allora si ha che:

1. esiste un unico punto fisso x_* di g in $[a, b]$;
2. la successione $x_{k+1} = g(x_k)$ generata dal metodo del punto fisso converge ad x_* per ogni punto iniziale $x_0 \in [a, b]$;
3. per ogni iterata del metodo si ha

$$|x_k - x_*| \leq \frac{L^k}{1 - L}|x_1 - x_0|.$$

Dimostrazione (Punto 2)

- Dall'ipotesi $g(x) \in [a,b]\ \forall\ x \in [a,b]$ segue che $x_k \in [a,b]$, $\forall\ k \geq 0$.
  Dimostriamo inoltre che la successione $\{x_k\}_{k\in\mathbb{N}}$ è di Cauchy, ossia:

$$\lim_{k\to\infty} |x_{k+p} - x_k| = 0,\ \forall p > 0.$$

Infatti per la disuguaglianza triangolare si ha

$$|x_k - x_{k+p}| = |x_k \pm x_{k+1} \pm \ldots \pm x_{k+p-1} - x_{k+p}| \leq \sum_{j=0}^{p-1} |x_{k+j} - x_{k+j+1}|.$$

Inoltre si ha

$$\begin{aligned} |x_{k+j} - x_{k+j+1}| &= |g(x_{k+j-1}) - g(x_{k+j})| \\ &\leq L|x_{k+j-1} - x_{k+j}| \leq \ldots \leq L^j |x_k - x_{k+1}|. \end{aligned}$$

Segue che

$$\begin{aligned} |x_k - x_{k+p}| &\leq \sum_{j=0}^{p-1} L^j |x_k - x_{k+1}| = \frac{1-L^p}{1-L}|x_k - x_{k+1}| \\ &\leq \frac{1}{1-L}|x_k - x_{k+1}| \leq \frac{L^k}{1-L}|x_0 - x_1| \overset{k\to\infty}{\longrightarrow} 0. \end{aligned}$$

Se $\{x_k\}_{k\in\mathbb{N}}$ è di Cauchy, allora converge ad un punto $x_* \in [a,b]$.

Dimostrazione (Punto 2)

- Per continuità di g, si ha

$$g(x_*) = g(\lim_{k\to\infty} x_k) = \lim_{k\to\infty} g(x_k) = \lim_{k\to\infty} x_{k+1} = x_*,$$

da cui segue che x_* è un punto fisso di g.

Dimostrazione (Punto 1)

- Se per assurdo supponiamo che g ammetta un altro punto fisso $y_* \in [a, b]$, ovvero se

$$g(y_*) = y_* \neq x_*,$$

allora dall'ipotesi di contrattività si avrebbe

$$L|x_* - y_*| \geq |g(x_*) - g(y_*)| = |x_* - y_*|, \quad \text{con } L < 1,$$

il che è assurdo.

Dimostrazione (Punto 3)

- Dalla dimostrazione del punto 2, abbiamo ottenuto che

$$|x_k - x_{k+p}| \leq \frac{L^k}{1-L}|x_0 - x_1|.$$

Prendendo il limite per $p \to \infty$ di entrambi i membri di tale disuguaglianza e osservando che il secondo membro non dipende da p, si ottiene

$$|x_k - x_*| \leq \frac{L^k}{1-L}|x_0 - x_1|.$$

Osservazioni

- Se viene a mancare una delle ipotesi del metodo, l'esistenza e/o l'unicità del punto fisso non sono più garantite.
- Una condizione sufficiente affinché una funzione differenziabile g sia contrattiva in $[a, b]$ è che $|g'(x)| \leq L < 1, \forall\, x \in [a, b]$.
- L'ipotesi di contrattività è analoga alla condizione necessaria e sufficiente per la convergenza di un metodo iterativo per sistemi lineari:

$$x^{(k+1)} = Gx^{(k)} + c \quad \text{converge se} \quad \rho(G) < 1$$
$$x_{k+1} = g(x_k) \quad \text{converge se} \quad |g'(x)| < 1.$$

- Dalla maggiorazione dell'errore

$$|x_k - x_*| \leq \frac{L^k}{1 - L}|x_0 - x_1|.$$

segue che

$$|x_k - x_*| = \mathcal{O}(L^k),$$

per cui tanto più L è piccolo, tanto più veloce sarà la convergenza (si noti l'analogia con $\rho(G)$).

Teorema

Se le ipotesi del teorema della mappa contrattiva sono soddisfatte ed inoltre di ha $g \in C^p([a,b])$, con $g^{(k)}(x_*) = 0$, $k = 1, \ldots, p-1$ e $g^{(p)}(x_*) \neq 0$, allora il metodo del punto fisso associato a g ha ordine p.

Dimostrazione

Dal teorema di Taylor e dalla definizione di punto fisso, per qualche ξ_k compreso tra x_* e x_k, si ha

$$\begin{aligned} x_{k+1} &= g(x_k) \\ &= g(x_*) + g'(x_*)(x_k - x_*) + \frac{1}{2}g''(x_*)(x_k - x_*)^2 + \ldots \\ &\ldots + \frac{1}{(p-1)!}g^{(p-1)}(x_*)(x_k - x_*)^{p-1} + \frac{1}{p!}g^{(p)}(\xi_k)(x_k - x_*)^p \\ &= x_* + \frac{1}{p!}g^{(p)}(\xi_k)(x_k - x_*)^p \end{aligned}$$

da cui si ricava

$$\frac{|x_{k+1} - x_*|}{|x_k - x_*|^p} = \frac{1}{p!}|g^{(p)}(\xi_k)| \overset{k\to\infty}{\longrightarrow} \frac{1}{p}|g^{(p)}(x_*)|$$

per continuità di g e perché ξ_k è compreso tra x_* e x_k.

- La convergenza quadratica del metodo di Newton si può ottenere come caso particolare del risultato precedente, dato che il metodo di Newton può essere visto come un metodo del punto fisso $x_{k+1} = g(x_k)$ con

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

$$g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)},$$

ed essendo x_* una radice di f segue che

$$g'(x_*) = \frac{f(x_*)f''(x_*)}{f'(x_*)} = 0.$$

2. Metodi per sistemi di equazioni non lineari

Definizione

- Una funzione scalare in n variabili $f : \mathbb{R}^n \to \mathbb{R}$ è una funzione che ad ogni vettore $x \in \mathbb{R}^n$ associa uno scalare $f(x) \in \mathbb{R}$.
- Una funzione vettoriale in n variabili $F : \mathbb{R}^n \to \mathbb{R}^n$ è una funzione che ad ogni vettore $x \in \mathbb{R}^n$ associa un vettore $F(x) \in \mathbb{R}^n$, dove

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad F(x) = \begin{pmatrix} f_1(x_1, \ldots, x_n) \\ f_2(x_1, \ldots, x_n) \\ \vdots \\ f_n(x_1, \ldots, x_n) \end{pmatrix},$$

con $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, n$ funzioni scalari di n variabili.

Esempi

- $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \sin(x)$ è una funzione scalare di 1 variabile;
- $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = \sin(x)\cos(y)$ è una funzione scalare di 2 variabili;
- $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x, y) = (xy, y)$ è una funzione vettoriale di 2 variabili.

Definizione (derivata prima per funzioni scalari di 1 variabile)

Dato $D \subseteq \mathbb{R}$ aperto, una funzione $f : D \to \mathbb{R}$ si dice derivabile in $x_0 \in D$ se esiste ed è finito il limite

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

e il valore di questo limite è la derivata prima di f in x_0.

Definizione (derivate parziali per funzioni scalari di 2 variabili)

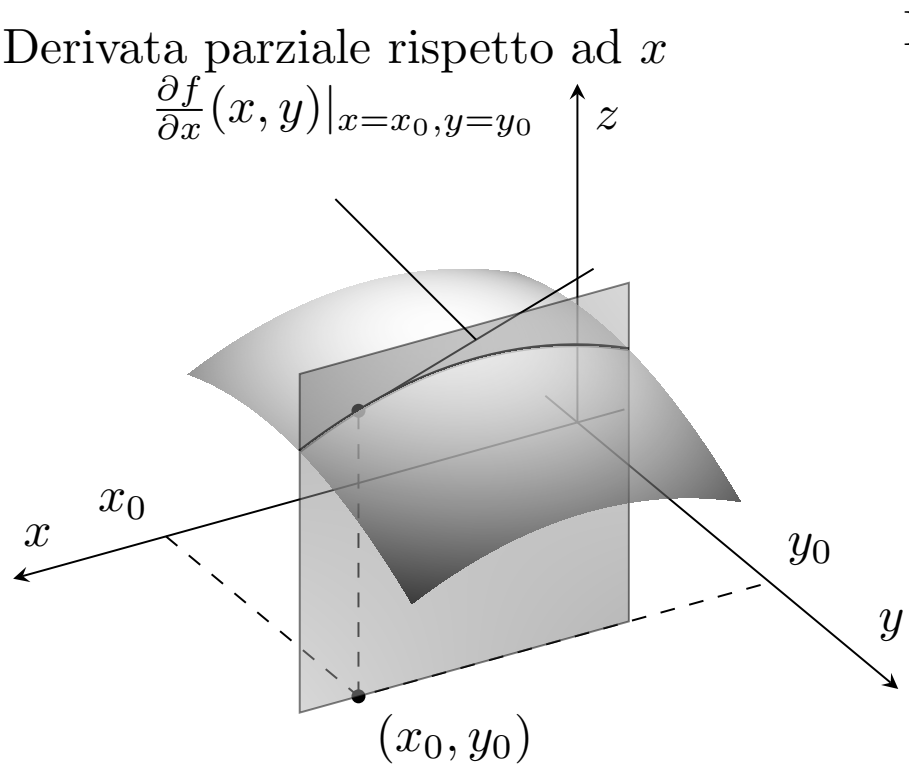
Dato $D \subseteq \mathbb{R}^2$ aperto e una funzione $f : D \to \mathbb{R}$, si definisce la derivata parziale di f rispetto ad x in un punto (x_0, y_0) il limite (se esiste) dato da

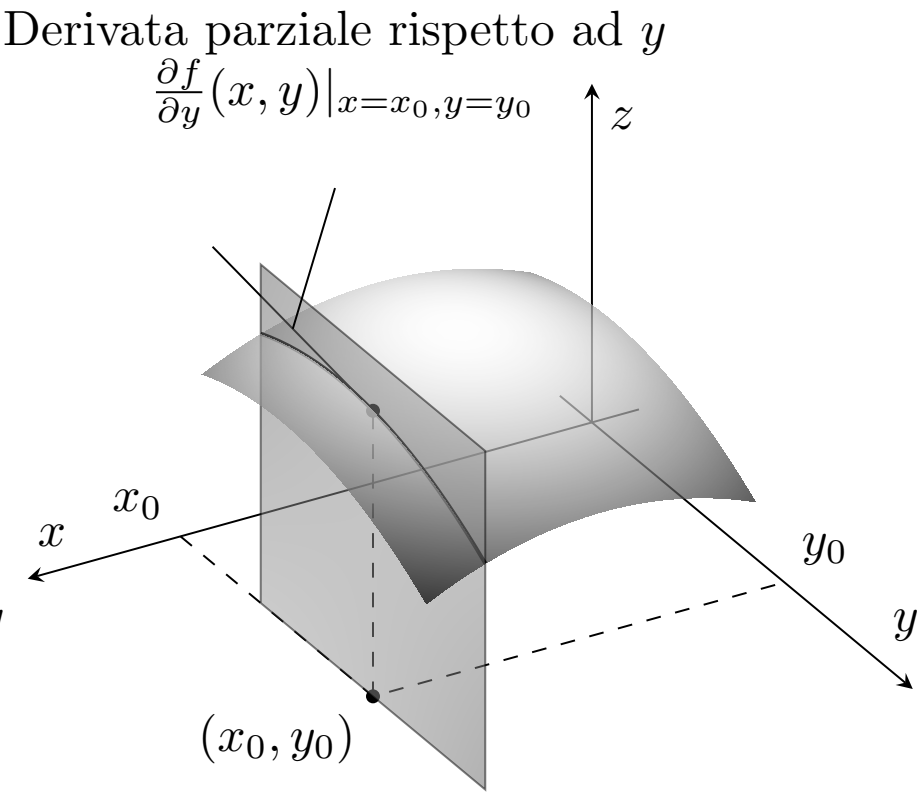
$$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h},$$

mentre la derivata parziale di f rispetto ad y in un punto (x_0, y_0) è il limite (se esiste) dato da

$$\frac{\partial f}{\partial y}(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0, y_0 + h) - f(x_0, y_0)}{h}.$$

- La derivata prima $f'(x)$ rappresenta la pendenza della retta tangente al grafico di f nel punto $(x, f(x))$.
- La derivata parziale $\frac{\partial f}{\partial x}$ (risp. $\frac{\partial f}{\partial y}$) rappresenta la pendenza della retta tangente alla curva ottenuta intersecando il grafico di f (una superficie di $\mathbb{R}^3$) con un piano passante per (x_0, y_0) e parallelo al piano $y = 0$ (risp. $x = 0$).





Definizione (derivate parziali per funzioni scalari di n variabili)

Dato $D \subseteq \mathbb{R}^n$ aperto e una funzione $f : D \to \mathbb{R}$, si definisce la derivata parziale di f rispetto ad x_j in un punto $(x_1, x_2, \ldots, x_n)$ il limite (se esiste) dato da

$$\frac{\partial f}{\partial x_j}(x_1, x_2, \ldots, x_n) = \lim_{h \to 0} \frac{f(x_1, x_2, \ldots, x_j + h, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{h}.$$

Il vettore che ha per componenti le derivate parziali di f, ovvero

$$\nabla f(x_1, \ldots, x_n) = \left(\frac{\partial f}{\partial x_1}(x_1, \ldots, x_n), \ldots, \frac{\partial f}{\partial x_n}(x_1, \ldots, x_n) \right)^T$$

è detto gradiente di f.

Proposizione (caso $n = 1$)

Sia $f : D \subseteq \mathbb{R} \to \mathbb{R}$ una funzione derivabile su D aperto.

- Se x_0 è un punto di minimo su D per f, ovvero $f(x_0) \leq f(x)$ per ogni $x \in D$, allora
$$f'(x_0) = 0.$$
- Se f è convessa, allora
$$x_0 \text{ è un punto di minimo per } f \text{ su } D \quad \Leftrightarrow \quad f'(x_0) = 0.$$

Proposizione (caso $n > 1$)

Sia $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ una funzione differenziabile su D aperto.

- Se x_0 è un punto di minimo su D per f, ovvero $f(x_0) \leq f(x)$ per ogni $x \in D$, allora
$$\nabla f(x_0) = 0.$$
- Se f è convessa, allora
$$x_0 \text{ è un punto di minimo per } f \text{ su } D \quad \Leftrightarrow \quad \nabla f(x_0) = 0.$$

Definizione

Sia $F : \mathbb{R}^n \to \mathbb{R}^n$. Si vuole trovare un vettore $x \in \mathbb{R}^n$ che soddisfa l'uguaglianza

$$F(x) = 0$$

il che equivale a risolvere n equazioni non lineari nelle incognite $x_1, \ldots, x_n$

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0 \\ f_2(x_1, \ldots, x_n) = 0 \\ \vdots \\ f_n(x_1, \ldots, x_n) = 0 \end{cases}.$$

L'uguaglianza $F(x) = 0$ prende il nome di sistema non lineare.

- Nel caso in cui $F(x) = Ax - b$, il sistema $F(x) = 0$ diventa lineare e può essere risolto facendo ricorso ai metodi già visti a lezione (fattorizzazione LU e QR, metodi iterativi di Jacobi e Gauss-Seidel, ecc.).

- Nel caso in cui $F(x)$ non sia lineare, è possibile risolvere $F(x) = 0$ estendendo il metodo di Newton e i metodi del punto fisso.

Dato $x \in \mathbb{R}^n$, si definisce la matrice Jacobiana $JF(x) \in \mathbb{R}^{n \times n}$ di F, contenente lungo le sue righe i gradienti delle funzioni $f_1, \ldots, f_n$ nel punto $x = (x_1, \ldots, x_n)^T$:

$$JF(x) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1}{\partial x_2(x)} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} & \cdots & \frac{\partial f_2(x)}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_n(x)}{\partial x_1} & \frac{\partial f_n(x)}{\partial x_n} & \cdots & \frac{\partial f_2(x)}{\partial x_n} \end{pmatrix}.$$

Il metodo di Newton può essere esteso da $f(x) = 0$ a $F(x) = 0$ come segue:

Metodo di Newton scalare

$$\begin{cases} f'(x_k)d_k = -f(x_k) \\ x_{k+1} = x_k + d_k \end{cases}$$

$\Rightarrow$

Metodo di Newton vettoriale

$$\begin{cases} JF(x^{(k)})d^{(k)} = -F(x^{(k)}) \\ x^{(k+1)} = x^{(k)} + d^{(k)}. \end{cases}$$

$\Rightarrow$ Nel caso $n-$dimensionale, il metodo di Newton richiede la risoluzione di un sistema lineare di ordine n ad ogni passo
$\Rightarrow$ Costoso!

Data una funzione $g : \mathbb{R}^n \to \mathbb{R}^n$ con

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad g(x) = \begin{pmatrix} g_1(x_1, \ldots, x_n) \\ g_2(x_1, \ldots, x_n) \\ \vdots \\ g_n(x_1, \ldots, x_n) \end{pmatrix}$$

il metodo del punto fisso associato a g per la risoluzione di un sistema non lineare $F(x) = 0$ è definito esattamente come nel caso scalare, ovvero:

$$\begin{cases} x^{(0)} \in \mathbb{R}^n \\ x^{(k+1)} = g(x^{(k)}), \quad k = 0, 1, \ldots \end{cases}.$$

Affinché il metodo converga ad un punto x^* tale che $F(x^*) = 0$, è necessario che g sia una contrazione, ovvero

$$\|g(x) - g(y)\| \leq L\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n,\ L \in (0, 1),$$

dove il valore assoluto è stato sostituito da una norma vettoriale $\|\cdot\|$.

3. Applicazione ai problemi di classificazione nel machine learning

Un algoritmo di apprendimento automatico (machine learning) è un algoritmo capace di "imparare" dai dati messi a disposizione.

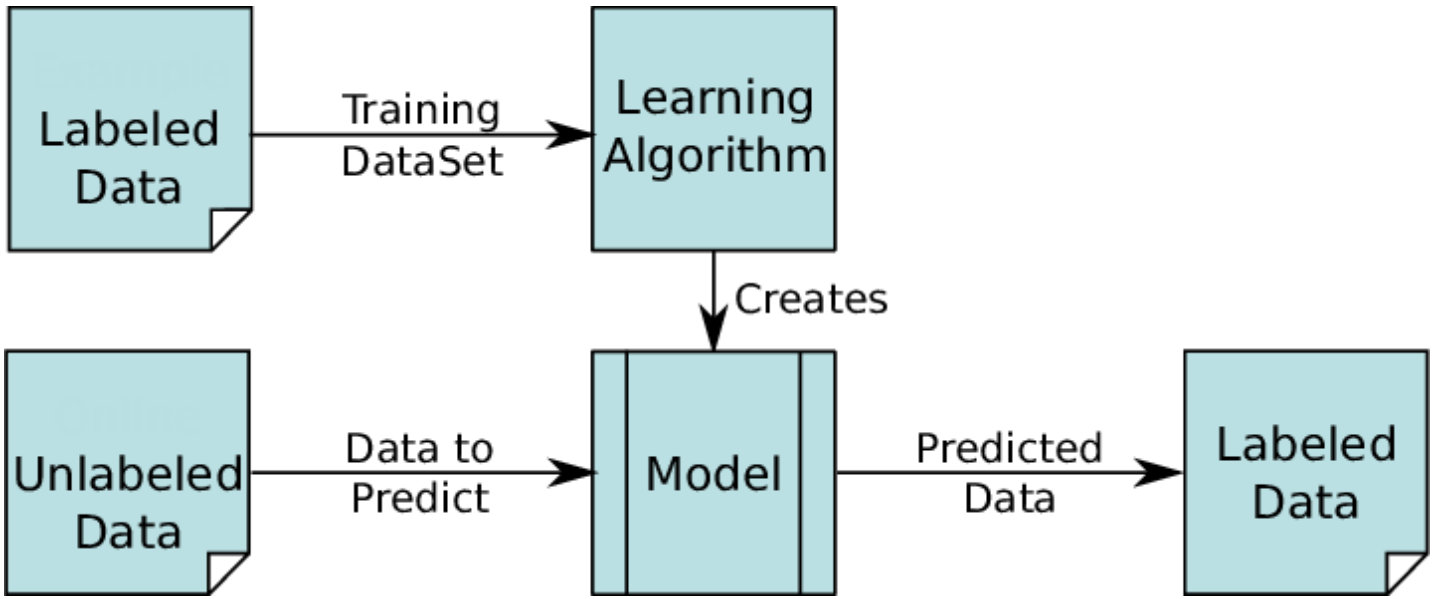
Mitchell, Machine Learning, McGraw-Hill, New York, 97, 1997.
Si dice che un programma di computer impara da un'esperienza E rispetto ad una classe di compiti (tasks) T e una misura di prestazione P, se la misura P relativamente a T migliora grazie all'esperienza E.

Esempio di problema di machine learning

- **Task**: **classificazione degli elementi di un insieme in uno o più gruppi**
- **Esperienza: insieme di dati di grandi dimensioni**
- **Misura: percentuale di elementi correttamente classificati**

Come fa un programma ad “imparare” dai dati messi a disposizione?

- *training set*: $\{(a_i, b_i)\}_{i=1,\dots,N}$
- *testing set*: $\{(a_i^{\text{test}}, b_i^{\text{test}})\}_{i=1,\dots,N_{\text{test}}}$
- a_i è detto esempio o vettore delle caratteristiche.
- b_i è detta etichetta associata all’esempio a_i.



Esempio di classificazione (1)

Dataset MNIST[1]

DATA	Training Size N	Test Size	Numero delle caratteristiche d
MNIST	60000	10000	784

Ciascun esempio consiste di 784 pixels "srotolati" dall'immagine 28×28 originale.

Classificazione di cifre:

stabilire quale cifra tra 0, 1,... 9 è rappresentata da una data immagine

[1] http://yann.lecun.com/exdb/mnist

Esempio di classificazione (2)

Dataset Mushrooms[1]

DATA	Training Size N	Test Size	Numero delle caratteristiche d
Mushrooms	5000	3124	112

Ciascun esempio consiste di 0 e 1; ognuna di queste cifre rappresenta una caratteristica del fungo dato (ad esempio se il cappello è marrone o no, liscio o no, ecc.)

.

Classificazione di funghi innocui e velenosi:
sicuro da mangiare o mortalmente velenoso?

[1] https://www.kaggle.com/uciml/mushroom-classification

Obiettivo

Determinare una funzione di predizione (modello) $h : \mathcal{A} \to \mathcal{B}$ tale che, dato un nuovo esempio $a \in \mathcal{A}$, il valore $h(a)$ offra un'accurata predizione della vera etichetta b associata all'input a.

Training

- Scegliere una funzione di predizione parametrizzata da un vettore $x \in \mathbb{R}^n$

$$h \in \mathcal{H} = \{h(\cdot; x) : x \in \mathbb{R}^n\}.$$

- Introdurre una funzione di loss $\ell : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ che, data una coppia input-output (a, b), restituisca l'errore (loss) $\ell(h(a; x), b)$ commesso nell'approssimare b con l'etichetta predetta $h(a; x)$.
- Dato un insieme di esempi $\{(a_i, b_i)\}_{i=1}^N$ (training set), $a_i \in \mathbb{R}^d$ (esempio), $b_i \in \mathbb{R}^p$ (etichetta), calcolare x_* come il punto di minimo della funzione $f : \mathbb{R}^n \to \mathbb{R}$ definita come

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} \underbrace{\ell(h(a_i; x), b_i)}_{\phi_i(x)}$$

Rischio empirico

Testing

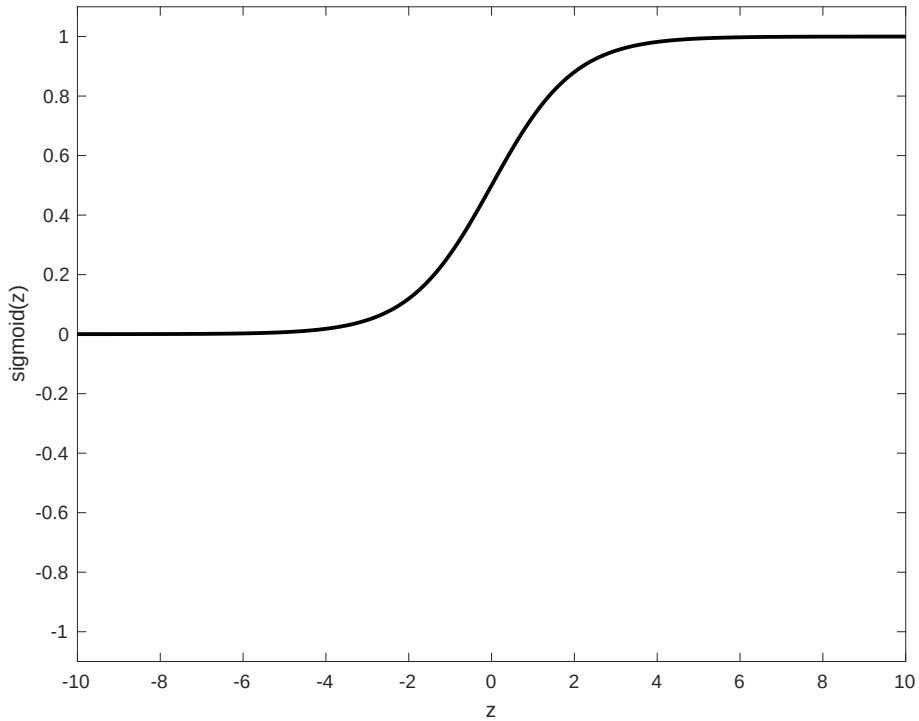
- Scegliere un testing set su cui valutare l'accuratezza della funzione di predizione $h(\cdot, x_*)$: quanti esempi del testing set vengono classificati correttamente?

Sia dato il training set $\{(a_i, b_i)\}_{i=1}^N$, $a_i \in \mathbb{R}^d$, $b_i \in \{-1, +1\}$.

- Si assume che la probabilità $P(b|a)$ che b sia l'etichetta di a sia data da

$$P(b|a) = \zeta(a, b; x) = \frac{1}{1 + e^{-ba^T x}}, \qquad \zeta(a, b; x) : \mathbb{R}^n \to (0, 1)$$

dove $x \in \mathbb{R}^d$ è un vettore di parametri da determinare.
La funzione $\zeta(z) = \frac{1}{1+e^{-z}}$ è detta sigmoide.



- Calcoliamo x in modo che sia massimizzato il prodotto delle probabilità degli eventi indipendenti "b_i è l'etichetta di a_i":

$$\max_{x \in \mathbb{R}^d} \prod_{i=1}^{N} P(b_i|a_i) = \max_{x \in \mathbb{R}^d} \prod_{i=1}^{N} \frac{1}{1 + e^{-b_i a_i^T x}}.$$

- Prendendo il logaritmo della funzione cambiata di segno:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{N} \sum_{i=1}^{N} \log(1 + e^{-b_i a_i^T x}).$$

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{N}\sum_{i=1}^{N} \log(1 + e^{-b_i a_i^T x})$$

- Dato x_* punto di minimo della funzione f, classifichiamo un nuovo elemento $\hat{a} \in \mathbb{R}^n$ come segue

$$h(\hat{a}; x_*) = \begin{cases} 1, & \text{se } P(1|\hat{a}) = \dfrac{1}{1+e^{-\hat{a}^T x_*}} \geq 0.5 \\ -1, & \text{se } P(1|\hat{a}) = \dfrac{1}{1+e^{-\hat{a}^T x_*}} < 0.5. \end{cases}$$

- Siccome f è convessa, segue che

$$x_* \text{ è punto di minimo di } f \quad \Leftrightarrow \quad \nabla f(x_*) = 0.$$

L'uguaglianza $\nabla f(x) = 0$ è a tutti gli effetti un sistema non lineare.

- Per risolvere $\nabla f(x) = 0$, possiamo applicare il metodo del punto fisso associato alla funzione $g(x) = x - \phi(x)F(x)$ con $F(x) = \nabla f(x)$, ottenendo

$$x^{(k+1)} = g(x^{(k)}) = x^{(k)} - \phi(x^{(k)})\nabla f(x^{(k)}), \quad k = 0, 1, \ldots$$

I metodi di questa forma sono detti metodi del gradiente e sono utilizzati diffusamente nei problemi di machine learning.

- Notiamo che il gradiente della funzione $f(x) = \frac{1}{N}\sum_{i=1}^{N} \log(1 + e^{-b_i a_i^T x})$ è Lipschitziano, ossia soddisfa alla seguente proprietà

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n,\ L > 0.$$

- Dunque, affinché la funzione g sia una contrazione, si può prendere ϕ come una funzione costante del tipo

$$\phi(x) = \alpha < \frac{1}{L},$$

ottenendo così il metodo

$$x^{(k+1)} = x^{(k)} - \alpha \nabla f(x^{(k)}), \quad k = 0, 1, \ldots$$

Tale metodo è

- meno costoso del metodo di Newton: richiede infatti la valutazione del gradiente ad ogni passo, ma non la risoluzione di un sistema lineare;
- più lento del metodo di Newton.