

# Statistica inferenziale- III

---

STATISTICA NUMERICA

A.Y. 2022-2023

# Outline

---

**1) Maximum Likelihood  
Estimation (MLE)**

STATISTICA NUMERICA, CAP. 6.2.4

**2) funzioni in più variabili**

STATISTICA NUMERICA, CAP. 6.2.4

**3) Ottimizzazione**

STATISTICA NUMERICA, CAP. 6.2.4

# 1) Maximum Likelihood Estimation (MLE)

---

STATISTICA NUMERICA, CAP. 6.2.4

# Stima dei parametri di una distribuzione con MLE

---

Supponiamo di volere stimare i parametri di una distribuzione.

Il metodo di **Massima Verosimiglianza o Maximum Likelihood Estimation (MLE)** è costituito da due fasi:

1. Costruire la funzione di verosimiglianza L
2. Massimizzare la funzione di verosimiglianza

Il metodo di massima verosimiglianza (o MLE) è una tecnica fondamentale nella statistica inferenziale per stimare i parametri di un modello statistico a partire dai dati osservati. L'idea centrale è trovare i valori dei parametri che rendono più probabile (o più verosimile) osservare i dati effettivamente raccolti.

# Funzione di verosimiglianza

La **verosimiglianza** non è la probabilità che il parametro sia un certo valore, ma la probabilità (o densità) di aver osservato quel dato condizionatamente a un certo valore del parametro.

Supponiamo che  $X_1, X_2, \dots, X_n$  sia un SRS(n) da una distribuzione con PDF o PMF  $f(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_m)$ . Se consideriamo **f** come funzione dipendente da **m parametri  $\theta_1, \theta_2, \dots, \theta_m$** , assegnati i valori osservati  $x_1, x_2, \dots, x_n$ , la funzione di verosimiglianza è:

Funzione di verosimiglianza

↓  
Formalmente definita come prodotto delle funzioni di densità valutata in ciascun punto del campione

$$L(\theta_1, \theta_2, \dots, \theta_m) = \prod_{i=1}^n f(x_i; \theta_1, \theta_2, \dots, \theta_m)$$

↓  
L'è una funzione dei parametri  $\Theta$ . I dati  $x_i$  sono fissati e noti.

PDF o PMF → questa funzione dipende da **m parametri incogniti  $\Theta$**

cosa si fa? → Fissiamo i valori osservati del campione  $(x_1, \dots, x_n)$  e consideriamo la funzione **f** non più come funzione della variabile **x**, ma come funzione dei parametri  $\Theta$ .

Perché la produttoria ( $\prod$ )?

↪ Le osservazioni sono indipendenti. La probabilità cangiante di osservare tutto il campione sotto ip. di indipendenza è la  $\prod$ . ricordo  $P(A \cap B) = P(A) \cdot P(B)$  se indipendenti

# Funzione di verosimiglianza: esempi

---

*Esempio 6.9 Si deve stimare il parametro di probabilità  $p$  di una distribuzione binomiale.*

La funzione di verosimiglianza in questo caso è:

$$L(p) = p^{\sum_i x_i} (1 - p)^{n - \sum_i x_i}$$

ottenuta come prodotto delle funzioni binomiali `binom(size = 1, prob = p)`.

Successivamente per stimare i parametri di una distribuzione  
bisogna massimizzare la funzione di verosimiglianza

# Funzione di verosimiglianza: esempi

**Esempio 6.8** Supponiamo ora che  $X_1, X_2, \dots, X_n$  siano un SRS(n) da una distribuzione normale con media  $\mu$  e deviazione standard  $\sigma$ . Vogliamo stimare i parametri  $\mu$  e  $\sigma$  con uno stimatore MLE.

La funzione di verosimiglianza è:

$$\begin{aligned} L(\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/\sigma^2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_2-\mu)^2/\sigma^2} \dots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/\sigma^2} \\ &= \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\sum_i (x_i-\mu)^2/\sigma^2}. \end{aligned}$$

Stimatore MLE

Successivamente per stimare i parametri di una distribuzione bisogna massimizzare la funzione di verosimiglianza

# MLE: fase II

---

- Non è detto che esista il massimo della funzione di verosimiglianza e anche se esiste non è detto che sia unico.
- Spesso è più semplice minimizzare l'opposto del logaritmo naturale di  $L(\theta)$ , cioè  $-\ln(L(\theta))$ , anzichè massimizzare  $L(\theta)$ . Poiché la funzione *logaritmo* è monotona, i due problemi hanno le stesse soluzioni (ricordiamo inoltre che  $\operatorname{argmin}_x f(x) = \operatorname{argmin}_x -f(x)$  per qualsiasi funzione  $f$ ).
  - Spesso non è possibile calcolare il minimo esattamente quindi si ricorre ad algoritmi numerici di ottimizzazione.


$$\operatorname{argmin} = \log(L(\theta_1, \dots, \theta_m))$$

Rapporto incrementale di una funzione

# 2) funzioni in più variabili

---

STATISTICA NUMERICA, CAP. 6.2.4

# Strumenti di calcolo differenziale in più variabili

Signifera che ogni punto  $P_0$  in  $A$ , esiste sempre un piccolo cerchio centrale in  $P_0$  in cui punti appartengono ancora ad  $A$ . Questo piccolo cerchio chiamato intorno di  $P_0$ , indicato come  $I(P_0, \delta)$   $\delta \rightarrow$  raggio.



Sia  $f(x, y)$  una funzione definita in un insieme aperto  $A \subset R^2$  e sia  $P_0 = (x_0, y_0)$  un punto di  $A$ .

Essendo  $A$  un aperto, esiste un intorno  $I(P_0, \delta) \subset A$ .

Preso un punto  $P(x, y) \in I(P_0, \delta)$ ,  $P \neq P_0$ , possiamo definire i due rapporti incrementali:

La derivata  $f'(x_0)$  misura il tasso di variazione istantaneo di  $f$  rispetto a  $x$  nel punto  $x_0$ . E' definita come limite del rapporto incrementale.  
Rispetto a quale variabile  
Stiamo misurando la variazione?

Rapporto incrementale parziale: isolano l'esigenza di una variabile mantenendo l'altra fissa.

$$\frac{f(x, y_0) - f(x_0, y_0)}{x - x_0}$$

Rapporto incrementale rispetto a  $x$

$$\frac{f(x_0, y) - f(x_0, y_0)}{y - y_0}$$

Rapporto incrementale rispetto a  $y$

Se esiste ed è finito il limite per  $x \rightarrow x_0$  del primo rapporto, si dice che la funzione  $f(x, y)$  è parzialmente derivabile rispetto a  $x$  nel punto  $P_0 = (x_0, y_0)$ .

# Le derivate parziali: definizione

---

Il valore di tale limite si chiama **derivata parziale rispetto a x nel punto  $P_0 = (x_0, y_0)$**  e si indica con uno dei seguenti simboli:

$$\frac{\partial f}{\partial x}(x_0, y_0), \quad f_x(x_0, y_0)$$

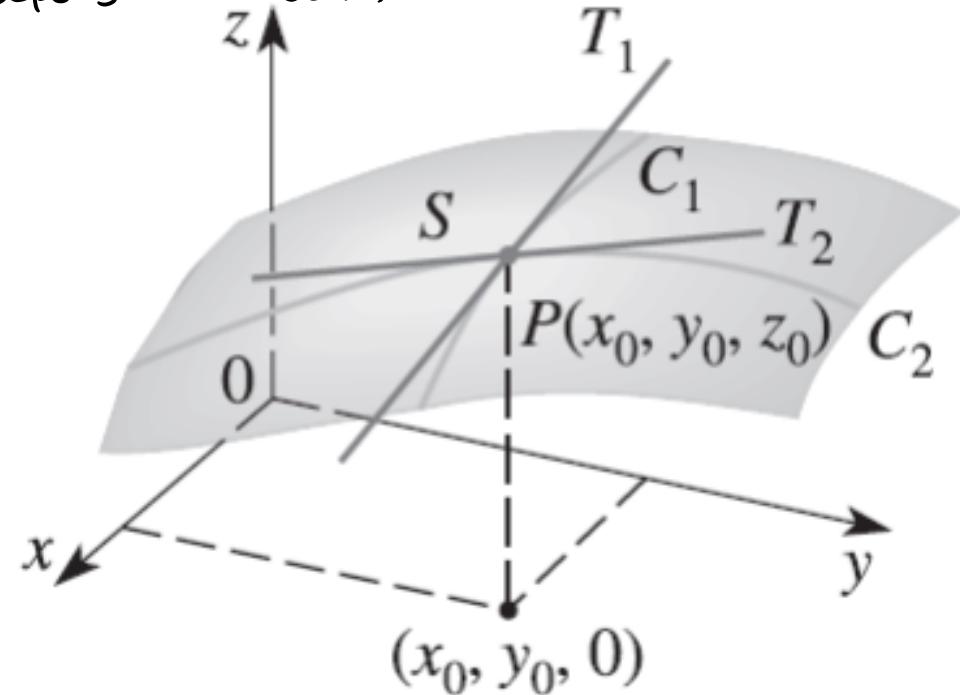
Analogamente se esiste il limite per  $y \rightarrow y_0$  del secondo rapporto, si dice che la funzione è parzialmente derivabile rispetto a y nel punto  $P_0 = (x_0, y_0)$ . Il valore di tale limite si chiama **derivata parziale rispetto a y della funzione nel punto  $(x_0, y_0)$**

$$\frac{\partial f}{\partial y}(x_0, y_0), \quad f_y(x_0, y_0)$$

Per le funzioni di una variabile la derivata è la pendenza (o coefficiente angolare) della retta tangente al grafico della funzione nel punto assegnato.

Le derivate parziali di una funzione di due variabili sono anch'esse legate alle pendenze di rette tangenti al grafico, ma di queste rette, ora, ce n'è più d'una.

$f_x(x_0, y_0)$  è la pendenza della retta tangente alla curva ottenuta intersecando la superficie  $z = f(x, y)$  con il piano  $x = x_0$  nel punto  $P_0$ .



## Le derivate parziali: significato geometrico

La derivata parziale  $\rightarrow$  ci fornisce la variazione istantanea di quota della funzione  $f(x,y)$  rispettivamente rispetto alla variabile  $x$  e alla  $y$ .

Consideriamo il paraboloide  $z = x^2 + y^2$

Le sue derivate parziali in un generico punto  $P(x,y)$  sono:

$$f_x(x,y) = 2x$$

$$f_y(x,y) = 2y$$

Se scegliamo  $(x_0, y_0) = (0,0)$  otteniamo

$$f_x(0,0) = 0$$

$$f_y(0,0) = 0$$

Se vogliamo invece calcolarlo nel punto  $(1,2, f(1,2))$

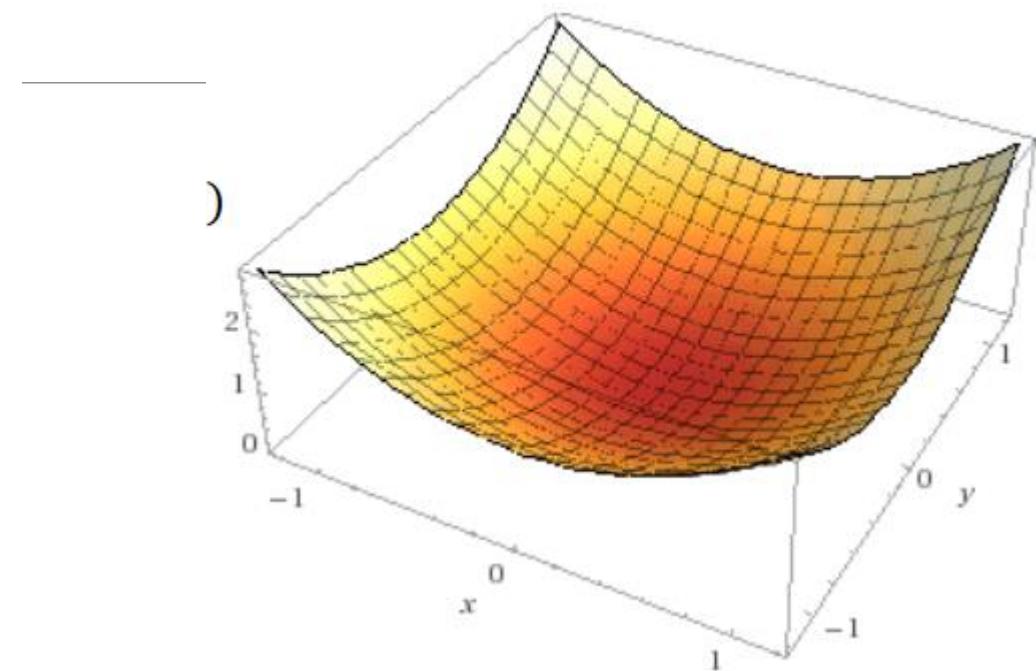
$$f(1,2) = 5$$

$$f_x(1,2) = 2$$

$$f_y(1,2) = 4$$

e risulta:

$$z = 5 + 2(x - 1) + 4(y - 2)$$



Le derivate parziali: significato geometrico

# Le derivate parziali seconde

---

Se  $f(x, y)$  è una funzione derivabile in un aperto  $A \subset R^2$ , le sue derivate parziali  $f_x(x, y)$  e  $f_y(x, y)$  sono funzioni di due variabili e possono essere a loro volta derivabili. Ad esempio, se  $f_x(x, y)$  è derivabile, è possibile calcolarne le derivate parziali rispetto ad  $x$  e ad  $y$ , che verranno indicate rispettivamente con i simboli equivalenti

$$\begin{array}{lll} \underline{f_{xx}(x, y)} & \frac{\partial^2 f}{\partial x^2}(x, y) & \frac{\partial}{\partial x} \frac{\partial f}{\partial y}(x, y) \\ \underbrace{f_{xy}(x, y)} & \frac{\partial^2 f}{\partial y \partial x}(x, y) & \frac{\partial}{\partial y} \frac{\partial f}{\partial x}(x, y) \end{array}$$

# Le derivate parziali seconde: definizione

---

$\circ \circ \circ \circ$

$\circ \circ \circ \circ$

Analogamente

Se  $f_y(x, y)$  è derivabile possiamo calcolare le derivate seconde e verranno indicate con i simboli equivalenti

Ci possono essere 4  
derivate seconde parziali:

$$\begin{array}{lll} f_{yx}(x, y) & \frac{\partial^2 f}{\partial x \partial y}(x, y) & \frac{\partial}{\partial x} \frac{\partial f}{\partial y}(x, y) \\ f_{yy}(x, y) & \frac{\partial^2 f}{\partial y^2}(x, y) & \frac{\partial}{\partial y} \frac{\partial f}{\partial y}(x, y) \end{array}$$

# Il gradiente di una funzione

---

Consideriamo una funzione  $f(x, y)$  definita su un insieme aperto  $A \subseteq \mathbb{R}^2$  e sia  $(x_0, y_0) \in A$ . Se esistono in  $(x_0, y_0)$  la derivata parziale rispetto ad  $x$  e la derivata parziale rispetto ad  $y$ , che indichiamo con  $f_x(x_0, y_0)$  e  $f_y(x_0, y_0)$  rispettivamente, allora è possibile costruire un vettore che ha per componenti le derivate parziali:

*sicuramente ha*

$$\nabla f(x_0, y_0) = (f_x(x_0, y_0), f_y(x_0, y_0))$$

$\frac{68}{6x}, \frac{68}{6y}$

Il vettore prende il nome di **gradiente della funzione**  $f$  valutato in  $(x_0, y_0)$ , o ancora  $\nabla f(x_0, y_0)$ .

In modo del tutto naturale la definizione di gradiente può essere estesa ad una funzione  $f : A \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ . Sarà sufficiente considerare il vettore che ha per componenti le  $n$  derivate parziali del primo ordine valutate nel punto  $\mathbf{x}_0 \in A$ .

In modo del tutto naturale la definizione di gradiente può essere estesa ad una funzione  $f : A \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ . Sarà sufficiente considerare il vettore che ha per componenti le  $n$  derivate parziali del primo ordine valutate nel punto  $\mathbf{x}_0 \in A$ .

Per funzioni di tre variabili  $f(x, y, z)$ , ad esempio, il gradiente sarà:

$$\nabla f(x_0, y_0, z_0) = (f_x(x_0, y_0, z_0), f_y(x_0, y_0, z_0), f_z(x_0, y_0, z_0))$$

# Il gradiente di una funzione

---

# Il gradiente di una funzione: esempi

**Facile:** determiniamo il gradiente della funzione  $f(x, y) = x^2 + y^2$  nel punto  $(x_0, y_0) = (1, 0)$ .

①

Calcoliamo la derivata parziale rispetto ad  $x$  e valutiamola nel punto  $(x_0, y_0) = (1, 0)$ :

②

$$f_x(x, y) = 2x \implies f_x(1, 0) = 2$$

Calcoliamo la derivata parziale rispetto ad  $y$  ed effettuiamo la valutazione:

$$f_y(x, y) = 2y \implies f_y(1, 0) = 0$$

Scriveremo quindi che il **gradiente della funzione valutato nel punto** è  $\nabla f(1, 0) = (2, 0)$ .

Nel punto  $(1, 0)$  il modo più rapido per salire sulla superficie è spostarsi lungo la direzione dell'asse  $x$  positivo. Perché la componente  $x$  del gradiente è positiva mentre  $y$  è 0.

Massima Decrescenza  $\rightarrow$  Il vettore opposto  $-\nabla f = (-2, 0)$  punta lungo l'asse  $x$  negativo. Significa che la discesa più rapida dal punto  $(1, 0)$  è spostarsi verso l'origine  $(0, 0)$

# Il gradiente di una funzione: esempi

---

**Medio:** calcoliamo il gradiente della funzione  $f(x, y) = \ln(xy^2)$  nel punto  $(1, -2)$ .

Partiamo con il calcolo delle derivate parziali:

$$\bullet f_x(x, y) = \frac{1}{xy^2} \cdot y^2 = \frac{1}{x} \implies f_x(1, -2) = 1$$

$$\bullet f_y(x, y) = \frac{1}{xy^2} \cdot 2xy = \frac{2}{y} \implies f_y(1, -2) = -1$$

# Significato geometrico del gradiente

---

Il gradiente di una funzione in un punto fornisce direzione e verso nei quali la funzione cresce più rapidamente.

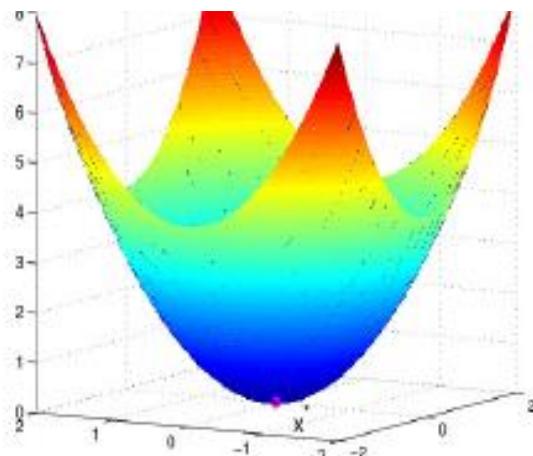
Nel verso opposto al gradiente avviene la massima decrescenza.

$$-\nabla f$$

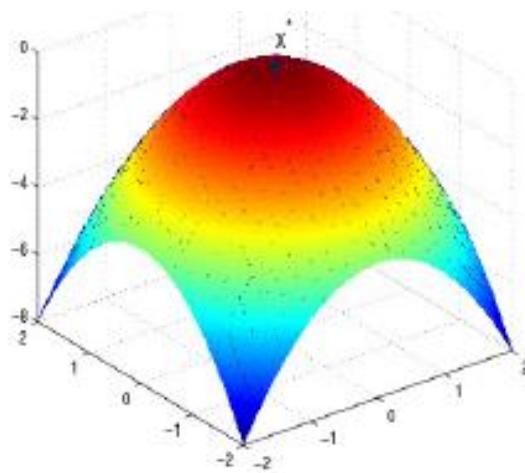
# 3) Ottimizzazione

---

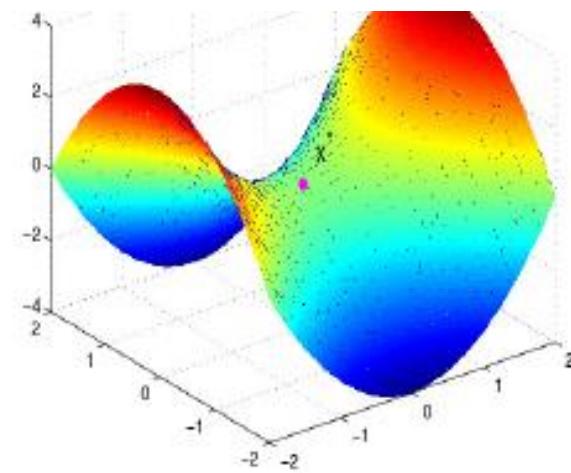
STATISTICA NUMERICA, CAP. 6.2.4



(a) Punto di minimo.



(b) Punto di massimo.



(c) Punto di sella.

# Ottimizzazione :concetti generali

# Massimizzare una funzione in più variabili

Gli algoritmi numerici per calcolare il massimo di una funzione in più variabili vengono detti algoritmi di ottimizzazione.

- ▶ L'ottimizzazione si occupa di risolvere problemi di estremo (massimo o minimo):

$$\min_x f(x) \quad (1)$$

dove  $x \in \mathbb{R}^n$  è un vettore reale di  $n \geq 1$  componenti e la funzione obiettivo  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  è una funzione regolare.

- ▶ Ottimizzazione deterministica (si conosce il modello) vs ottimizzazione stocastica (il modello è solo basato sulla probabilità)
- ▶ Ottimizzazione locale (calcola soluzioni locali del problema) vs ottimizzazione globale (calcola soluzioni globali del problema).

La funzione obiettivo serve a misurare le prestazioni di un sistema o la qualità di una soluzione. Il suo scopo è fornire un criterio numerico, un punteggio per confrontare diverse alternative e dire oggettivamente quale sia "migliore" o "peggiore".

# Ottimizzazione :concetti generali

Consideriamo il problema di **ottimizzazione non vincolata**:

$$\min_x f(x) \quad (2)$$

dove  $x \in \mathbb{R}^n$  è un vettore reale di  $n \geq 1$  componenti e la **funzione obiettivo**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  è una funzione regolare.

Un vettore  $\underline{x^*}$  è un punto di **minimo locale** di  $f(x)$  se esiste un  $\epsilon > 0$  tale

$$\underline{f(x^*) \leq f(x)} \quad \text{per ogni } \underline{x \text{ tale che } \|x - x^*\| < \epsilon}. \quad (3)$$

# Ottimizzazione :concetti generali

---

Un vettore  $x^*$  è un punto di minimo globale di  $f(x)$  se

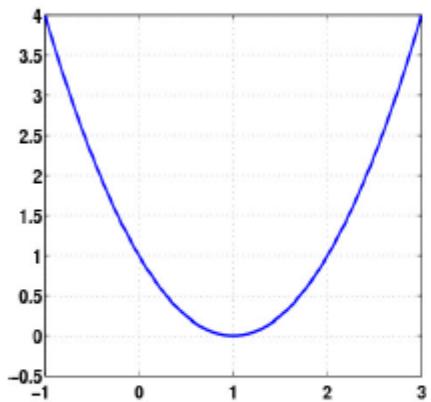
$$f(x^*) \leq f(x) \quad \text{per ogni } x \in \mathbb{R}^n. \quad (4)$$

Analogamente,  $x^*$  è un punto di minimo globale in senso stretto di  $f(x)$  se

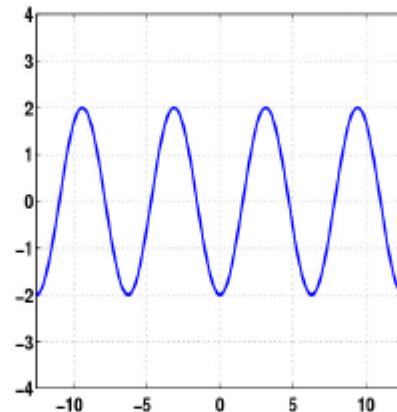
$$f(x^*) < f(x) \quad \text{per ogni } x \in \mathbb{R}^n, \quad x \neq x^*. \quad (5)$$

# Ottimizzazione :concetti generali

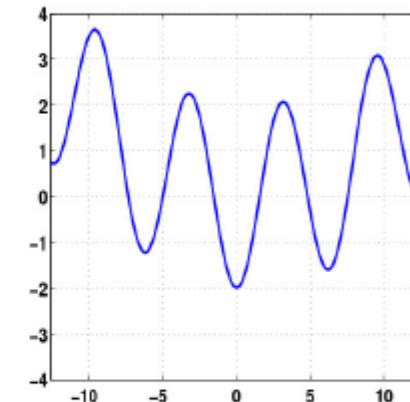
Una funzione  $f(x)$  può avere un unico punto di minimo locale (quindi anche globale), oppure può non avere nè minimi locali nè globali, può avere sia minimi locali che globali... (figure 1).



(a)  $y = (x - x^*)^2$  un  
unico punto di minimo



(b)  $y = -2 \cos(x - x^*)$   
molti punti di minimo  
globale.



(c)  $y = 0.015(x - x^*)^2 - 2 \cos(x - x^*)$  un punto  
di minimo globale e molti  
punti di minimo locale.

# Ottimizzazione :concetti generali

Teorema (Condizioni necessarie del primo ordine)

Se  $x^*$  è un punto di minimo locale e  $f$  è differenziabile con continuità in un intorno aperto di  $x^*$ , allora  $\nabla f(x^*) = 0$ .

↗ tutte le derivate  
prime parziali di  $f$

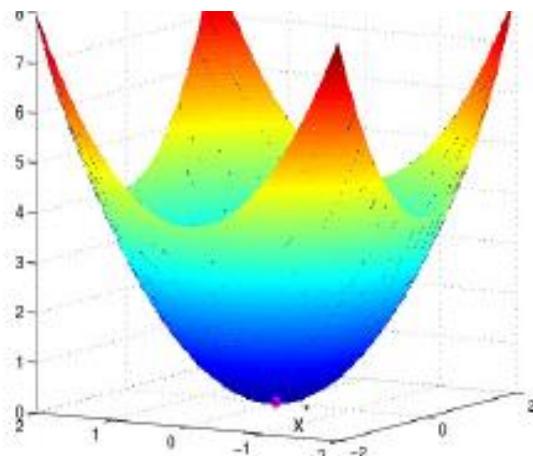
Un punto  $x^*$  tale che  $\nabla f(x^*) = 0$  è detto **punto stazionario**. Dal teorema segue che

$$x^* \text{ punto di minimo} \Rightarrow x^* \text{ punto stazionario}$$

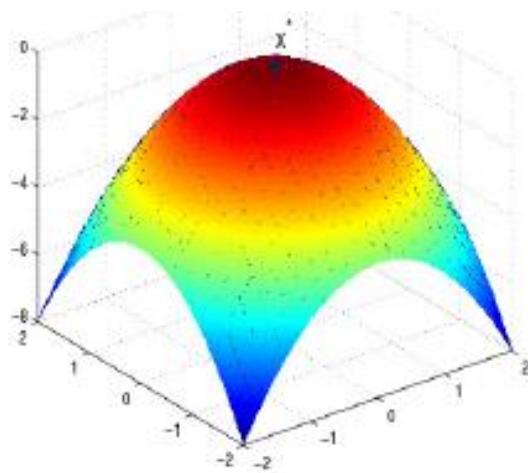
$x^*$  punto di minimo  $\Rightarrow \nabla f(x^*) = 0$



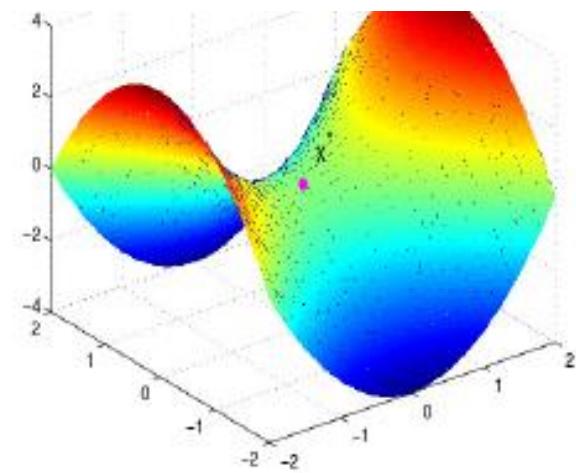
Condizione necessaria  
non sufficiente



(a) Punto di minimo.



(b) Punto di massimo.



(c) Punto di sella.

# Ottimizzazione :concetti generali

Sia  $f: \mathbb{R}^n \rightarrow \mathbb{R}$

La condizione  $\nabla f(x^*)$  è condizione necessaria affinchè  $x^*$  sia un punto di minimo locale; tale condizione non è però sufficiente poichè un punto stazionario può essere un punto di minimo locale, un punto di massimo locale o un punto di sella (figura 2). I punti di minimo locale possono essere distinti dagli altri punti stazionari esaminando le derivate seconde.

# Ottimizzazione: concetti generali

---

# Algoritmi di ottimizzazione classici

---

Gli algoritmi di ottimizzazione sono algoritmi iterativi.

- ▶ Gli algoritmi iterativi calcolano una sequenza di iterati  $x_1, x_2, \dots$ , a partire da un iterato iniziale  $x_0$  assegnato, secondo una legge del tipo:

$$x_{k+1} = G(x_k).$$

- ▶ La successione  $\{x_k\}$  deve avere proprietà di **convergenza** alla soluzione esatta  $x^*$ :

$$\lim_{k \rightarrow \infty} x_k = x^*$$

# Algoritmi di ottimizzazione classici

---

- Gli algoritmi iterativi per la minimizzazione di una funzione in generale NON convergono al minimo globale, ma ad un minimo locale.
- Tutti necessitano di un **iterato iniziale**, cioè di un vettore  $x_0$  che inneschi il metodo iterativo.
- L'iterato iniziale influenza il minimo locale a cui converge il metodo. Quindi se si ha una stima della soluzione desiderata, si deve scegliere l' iterato iniziale utilizzando tale stima.

# Metodi di discesa

Criteri di arresto

$$x_0 \text{ input } x_0, x_1, \dots, x_k, x_{k+1} \rightarrow x^*$$

$$\left[ \begin{array}{l} k=0 \text{ finche' criterio vero} \\ x_{k+1} = G \\ k=k+1 \end{array} \right]$$

$$x_k \in \mathbb{R}^n$$

- Si consideri il problema della minimizzazione non vincolata di una funzione  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  differenziabile con continuità.
- I metodi con ricerca in linea sono metodi iterativi che, a partire da un iterato iniziale  $x_0 \in \mathbb{R}^n$ , generano una successione di vettori

$$x_0, x_1, x_2, \dots$$

definiti dall'iterazione

$$x_{k+1} = x_k + \underline{\alpha_k p_k} \quad (2)$$

dove il vettore  $p_k$  è una **direzione di ricerca** e lo scalare  $\alpha_k$  è un parametro positivo chiamato **lunghezza del passo** (step-length) che indica la distanza di cui ci si deve muovere lungo la direzione  $p_k$ .

# Metodi di discesa

---

- Il vettore  $p_k$  ed il parametro  $\alpha_k$  sono scelti in modo da garantire la decrescita di  $f(x)$  ad ogni iterazione:

$$\nabla f(x_{k+1}) < f(x_k), \quad k = 0, 1, \dots$$

Definizione

Il vettore  $p$  è una direzione di discesa di  $f$  in  $x$  se esiste un  $\bar{\alpha} > 0$  tale che

$$f(x + \alpha p) < f(x) \text{ per ogni } \alpha \in (0, \bar{\alpha}]$$

- La direzione dell'antigradiente  $\underline{p} = -\nabla f(x)$  è sempre una direzione di discesa

# Metodi di discesa

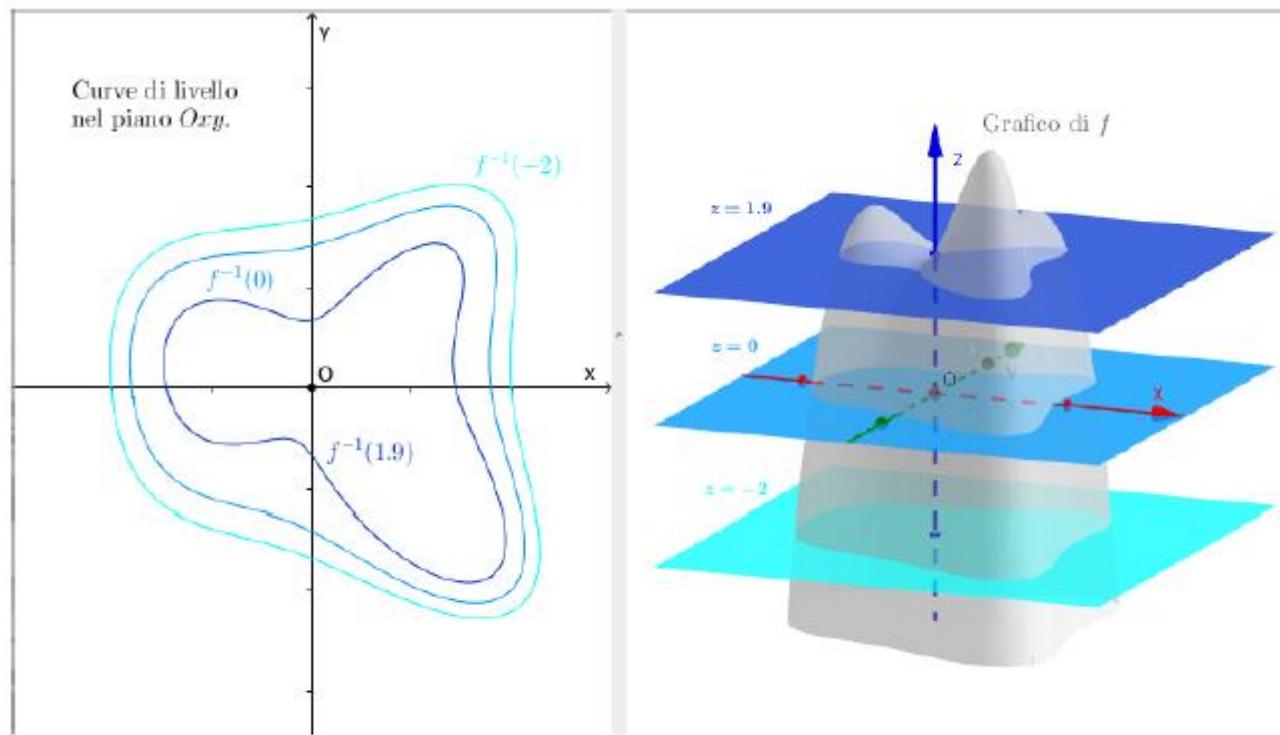
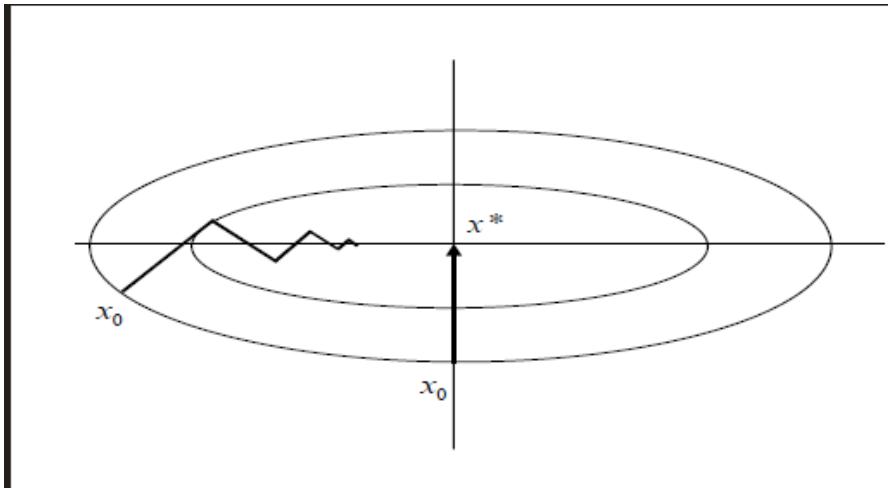


Figura 2.4: Esempio di curve di livello

# Metodo del gradiente

---



Algoritmo del metodo del gradiente

Input:  $f, x_0, \alpha$

$k=0$

Ripeti fino a convergenza

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

$k=k+1$

Algoritmi di Ottimizzazione Vincolata → le variabili decisionali devono soddisfare una serie di restrizioni (vincoli). Questi vincoli definiscono una regione ammessa, un sottoinsieme di  $\mathbb{R}^n$  all'interno del quale dobbiamo cercare la soluzione ottima.

# Metodi di discesa

Algoritmi di Ottimizzazione Non vincolata → Nell'ottimizzazione non vincolata, non ci sono restrizioni sui valori che le variabili decisionali possono assumere. L'obiettivo è trovare il punto di minimo o massimo della funzione obiettivo esplorando liberamente tutto lo spazio  $\mathbb{R}^n$ .

- Negli algoritmi di ottimizzazione esistono diversi criteri per far terminare le iterazioni e che possono indicare il raggiungimento di una soluzione oppure il fallimento dell'algoritmo.
- Negli algoritmi di ottimizzazione non vincolata per criterio di arresto si intende il criterio che dovrebbe indicare il raggiungimento con successo di un punto stazionario con la tolleranza specificata dall'utente.
- Dal punto di vista teorico, il criterio di arresto di un algoritmo che genera la successione  $x_k$  dovrebbe essere la condizione

$$\|\nabla f(x_k)\| \leq \epsilon, \quad \epsilon > 0 \quad (1)$$

# Funzioni Python

---

La libreria Python che contiene funzioni per ottimizzare funzioni in piu variabili e'  
[scipy.optimize](#)

Noi useremo in particolare la funzione [scipy.optimize.minimize](#).

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html>

Codici di esempio:

- 1) [mle\\_binom.py](#)
- 2) [mle\\_norm.py](#)