Statistica inferenziale- III

STATISTICA NUMERICA

A.Y. 2022-2023

Outline

1) Maximum Likelihhod 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 Likelihhod 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

$\theta = (\theta_1, \theta_2 .. \theta_m)$

Funzione di verosimiglianza

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

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

funzione verosimiglianza

$X = (x_1, x_2 .. x_n)$

$L : R^m \longrightarrow R$

$m = 1 \rightarrow L : R \rightarrow R$

$$L(\theta_1, \theta_2, \dots \theta_m) = \underline{f_X(x_1; \theta_1, \theta_2 \dots \theta_m)}$$

Poisson

$$f_X(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

retta

$$f(x; m, q) = \underline{mx + q}$$

Funzione di verosimiglianza: esempi

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

$n \to$ noto

La funzione di verosimiglianza in questo caso è:

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

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

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\left[\binom{n}{x_1} p^{x_1}(1-p)^{n-x_1}\right] \cdot \left[\binom{n}{x_2} p^{x_2}(1-p)^{n-x_2}\right] \ldots$$

SRS(μ) da distribuzione normale

Funzione di verosimiglianza: esempi

$a^{x_1} \cdot a^{x_2} = a^{x_1+x_2}$

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

La funzione di verosimiglianza è:

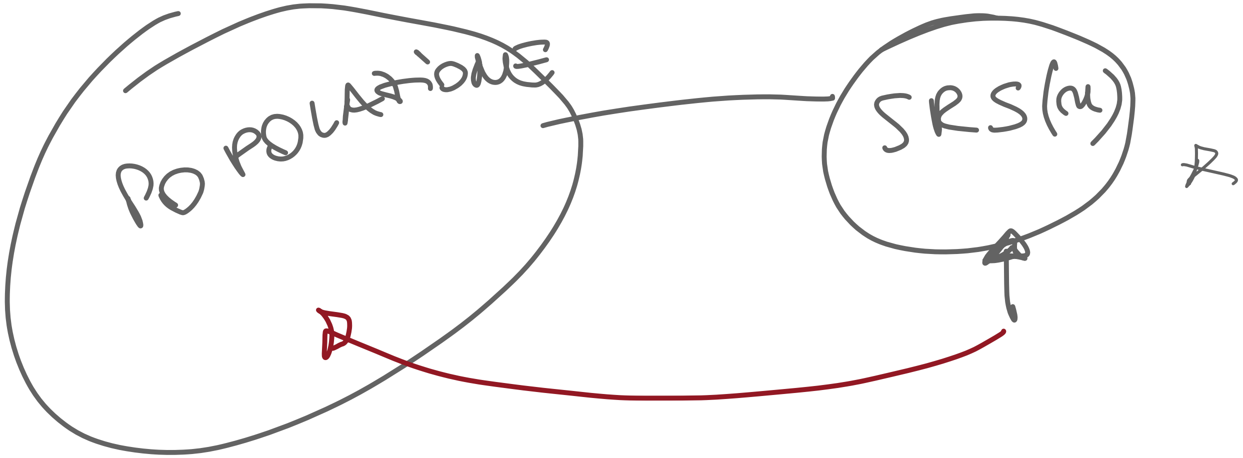
$$
\begin{aligned}
L(\mu, \sigma^2) &= \boxed{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/\sigma^2}}\ \boxed{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_2-\mu)^2/\sigma^2}} \cdots \boxed{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/\sigma^2}} \\
L(\theta_1, \theta_2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \boxed{e^{-\sum_i (x_i-\mu)^2/\sigma^2}}.
\end{aligned}
$$

$L : R^2 \to R$

$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\frac{n}{2}}$

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

STIMA dei PARAMETRI



- media $= \mu$
- varianza $= \sigma^2$
- parametri della distribuzione di probabilità

MAXIMUM LIKEHOOD
(massima verosimiglianza)

1). Calcolare di funzione di likelihood ←

2). calcolare il massimo della funzione

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 $argmin_x f(x) = argmin_x - f(x)$ per qualsiasi funzione f).

- Spesso non e' possibile calcolare il minimo esattamente quindi si ricorre ad algoritmi numerici di ottimizzazione.

$$\boxed{\text{argmin} - \log(L(\theta_1, \theta_2 .. \theta_m))}$$

2) funzioni in più variabili

STATISTICA NUMERICA, CAP. 6.2.4

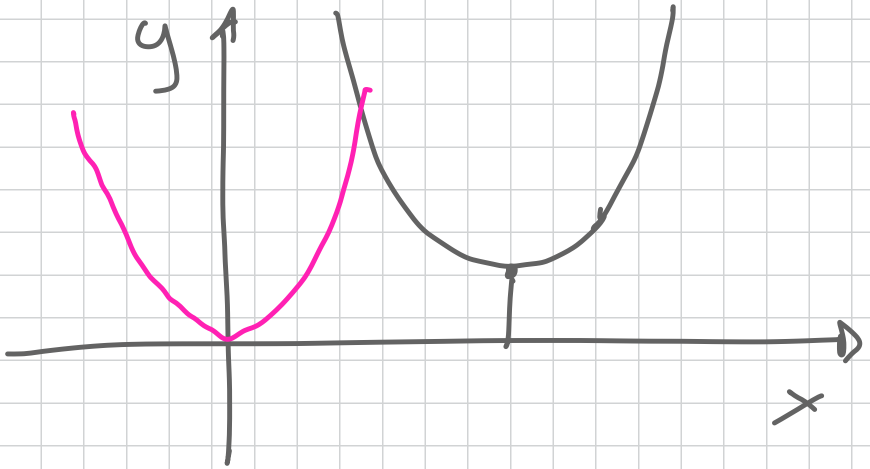
$f: \mathbb{R} \to \mathbb{R}$

$[a,b] \subset \mathbb{R} \to \mathbb{R}$

$f(x) = y = x^2$

e^x

$\log x$



$f: \mathbb{R}^2 \to \mathbb{R}$ $\quad f: [a,b] \times [c,d] \to \mathbb{R}$

$f: (x_1, x_2) \to f(x_1, x_2) = z$

esempi

$f(x_1, x_2) = x_1^2 + x_2^2$

$f(x_1, x_2) = \sqrt{x_1 + x_2}$

Derivata $f: R \to R$

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

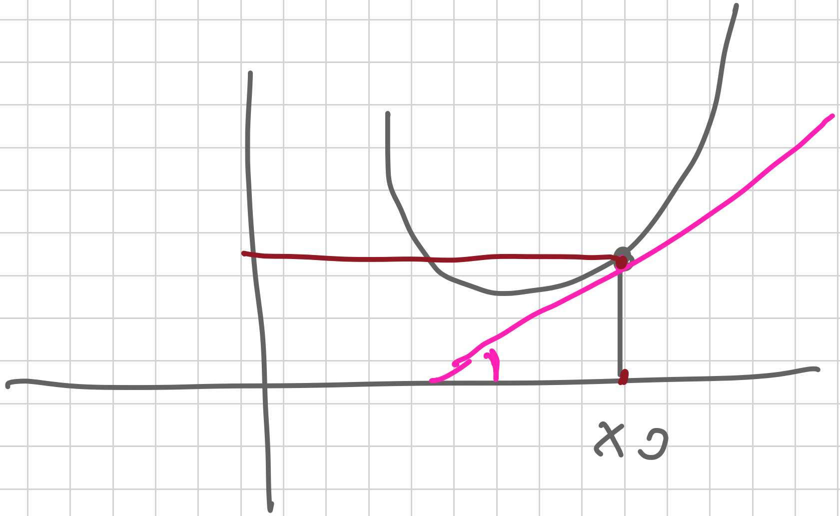
rapporto incrementale

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

$x - x_0 = h$

$\Rightarrow f'(x): R \to R$

$\Rightarrow f'(x_0) \in R$



Strumenti di calcolo differenziale in piu variabili

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:

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

Rapporto incrementale rispetto a x

e

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

Rapporto incrementale rispetto a y

Se esiste ed è finito il limite per $x \to x_0$ 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

$\frac{\partial f}{\partial x}: \mathbb{R}^2 \to \mathbb{R}$

$\frac{\partial f}{\partial y}: \mathbb{R}^2 \to \mathbb{R}$

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:

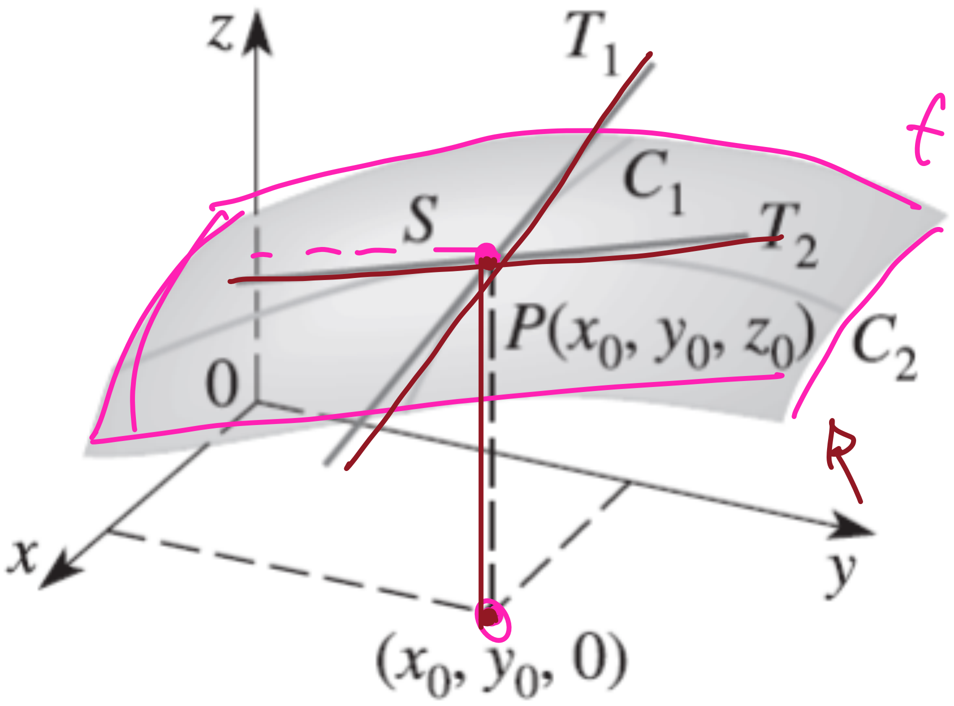
$$\boxed{\frac{\partial f}{\partial x}(x_0, y_0),} \qquad \boxed{f_x(x_0, y_0)}$$

Analogamente se esiste il limite per $y \to y_0$ 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)

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

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.



Le derivate parziali: significato geometrico

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

y

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

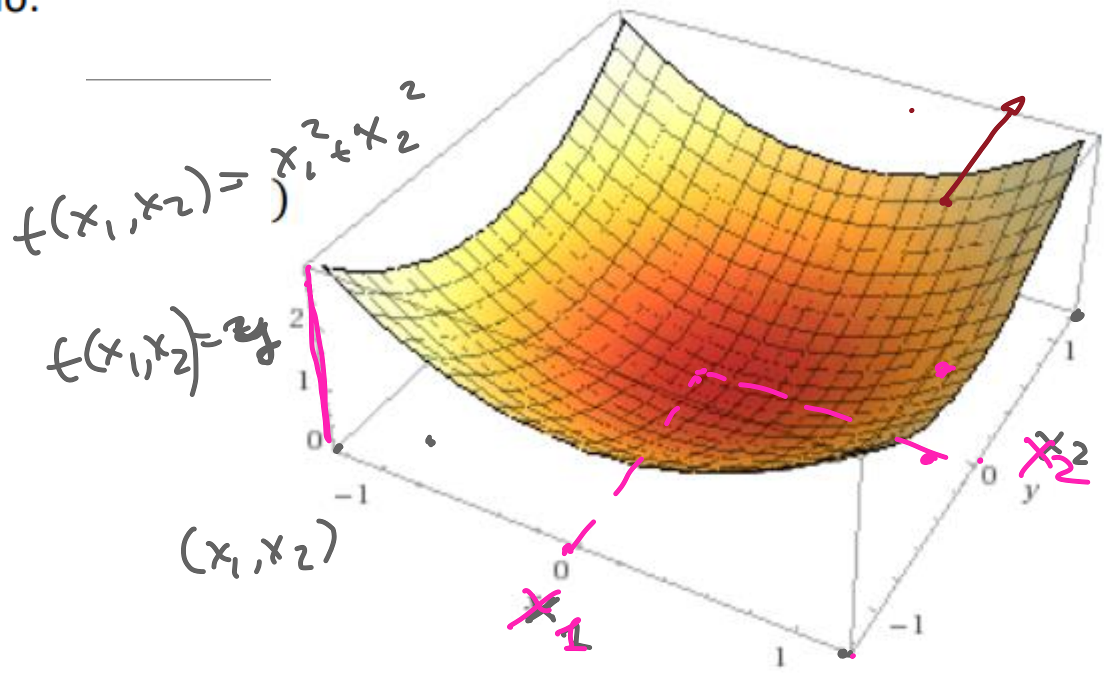
$$f_x(0,0) = 0$$
$$f_y(0,0) = 0$$

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

ESEMPI

$$f(x_1, x_2) = \underbrace{x_1^2} + x_2^2$$

$$\frac{\partial f}{\partial x_1} = 2x_1 + 0 = 2x_1.$$

$$\frac{\partial f}{\partial x_2} = 0 + 2x_2 = 2x_2.$$

2) $f(x_1, x_2) = \underbrace{3x_1x_2} + e^{x_1} - x_2^2$

$$\frac{\partial f}{\partial x_1} = 3x_2 + e^{x_1} \quad : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

$$\frac{\partial f}{\partial x_2} = 3x_1 - 2x_2 \quad : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

3) $f(x_1, x_2, x_3) = \underbrace{x_1 x_2^2} + x_3^3 + 5 \sin x_1 \cos x_3$

$$\frac{\partial f}{\partial x_1} = x_2^2 + 5\cos x_1 \cdot \cos x_3$$

$$\frac{\partial f}{\partial x_2} = 2x_2 \cdot x_1$$

$$\frac{\partial f}{\partial x_3} = 3x_3^2 - 5\sin x_1 \sin x_3$$

$$: \mathbb{R}^3 \longrightarrow \mathbb{R}$$

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

$$f_{xx}(x, y) \qquad \boxed{\frac{\partial^2 f}{\partial x^2}(x,, y)} \qquad \frac{\partial}{\partial x}\frac{\partial f}{\partial y}(x, y)$$

$$f_{xy}(x, y) \qquad \frac{\partial^2 f}{\partial y \partial x}(x,, y) \qquad \frac{\partial}{\partial y}\frac{\partial f}{\partial x}(x, y)$$

$f : \mathbb{R}^2 \to \mathbb{R}$

$\dfrac{\partial f}{\partial x_1}$ $\quad$ $\dfrac{\partial f}{\partial x_2}$ $\longrightarrow$ derivate parziali prime

① $\dfrac{\partial^2 f}{\partial x_1 \partial x_1} = \dfrac{\partial^2 f}{\partial x_1^2}$

② $\dfrac{\partial^2 f}{\partial x_1 \partial x_2}$

③ $\dfrac{\partial^2 f}{\partial x_2 \partial x_1}$

④ $\dfrac{\partial^2 f}{\partial x_2 \partial x_2} = \dfrac{\partial^2 f}{\partial x_2^2}$

$$f(x_1, x_2) = 3x_1 x_2^2 - x_1^3 x_2^3$$

$$\frac{\partial f}{\partial x_1} = \underbrace{3x_2^2}\ \underbrace{-3x_1^2 \cdot x_2^3} \quad : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

$$\frac{\partial f}{\partial x_2} = 6x_2 \cdot x_1 - 3x_1^3 x_2^2 \quad : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

$$\frac{\partial f}{\partial x_1}\begin{pmatrix} \overset{x_1}{0}, & \overset{x_2}{1} \end{pmatrix} = 3 - 0 = \boxed{3}$$

$$\frac{\partial f}{\partial x_2}(0, 1) = 0 - 0 = \boxed{0}$$

$$\frac{\partial^2 f}{\partial x_1^2} = -6x_1 \cdot x_2^3$$

$$\boxed{\frac{\partial^2 f}{\partial x_1 \partial x_2} = 6x_2 - 3x_1^2 \cdot 3x_2^2 = 6x_2 - 9x_1^2 x_2^2}$$

$$\boxed{\frac{\partial^2 f}{\partial x_2 \partial x_1} = 6x_2 - 9x_1^2 x_2^2}$$

$$\frac{\partial^2 f}{\partial x_2^2} = +6x_1 - 3x_1^3 \cdot 2x_2 = +6x_1 - 6x_1^3 x_2$$

Le derivate parziali seconde: definizione

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

$$f_{yx}(x,y) \quad \frac{\partial^2 f}{\partial x \partial y}(x,,y) \quad \frac{\partial}{\partial x}\frac{\partial f}{\partial y}(x,y)$$

$$f_{yy}(x,y) \quad \frac{\partial^2 f}{\partial y^2}(x,,y) \quad \frac{\partial}{\partial y}\frac{\partial f}{\partial y}(x,y)$$

$\boxed{f: R^2 \to R}$, $\boxed{\nabla f : R^2 \to R^2}$

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:

$$\boxed{\nabla f(x_0, y_0) = (f_x(x_0, y_0), f_y(x_0, y_0))} \leftarrow \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$$

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 $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$. Sarà sufficiente considerare il vettore che ha per componenti le n derivate parziali del primo ordine valutate nel punto $\mathbf{x}_0 \in A$.

$$f : \mathbb{R}^2 \to \mathbb{R}$$

$$\nabla f : \mathbb{R}^2 \to \mathbb{R}^2 \longrightarrow$$

$$\nabla f(x_1, x_2) = \left(\frac{\partial f}{\partial x_1}(x_1, x_2), \frac{\partial f}{\partial x_2}(x_1, x_2) \right) \in \mathbb{R}^2$$

$$P(P_1, P_2)$$

$$\nabla f(P_1, P_2) = \left(\frac{\partial f}{\partial x_1}(P_1, P_2), \frac{\partial f}{\partial x_2}(P_1, P_2) \right) \in \mathbb{R}^2$$

$$f(x_1, x_2) = x_1^2 + x_2^2$$

$$\nabla f(x_1, x_2) = (2x_1, 2x_2) \quad : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$$

$$\nabla f(\underline{0,0}) = (2 \cdot 0, 2 \cdot 0) = (\underline{0,0})$$

$$\nabla f(2,3) = (2 \cdot 2, 2 \cdot 3) = (4,6) \in \mathbb{R}^2$$

$$f : \mathbb{R}^n \longrightarrow \mathbb{R} \qquad n \geqslant 2$$

$$\nabla f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots \frac{\partial f}{\partial_n} \right)$$

In modo del tutto naturale la definizione di gradiente può essere estesa ad una funzione $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$. 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 $(x_0, y_0) = (1, 0)$.

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

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

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 $\nabla f(1, 0) = (2, 0)$.

Il gradiente di una funzione: esempi

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

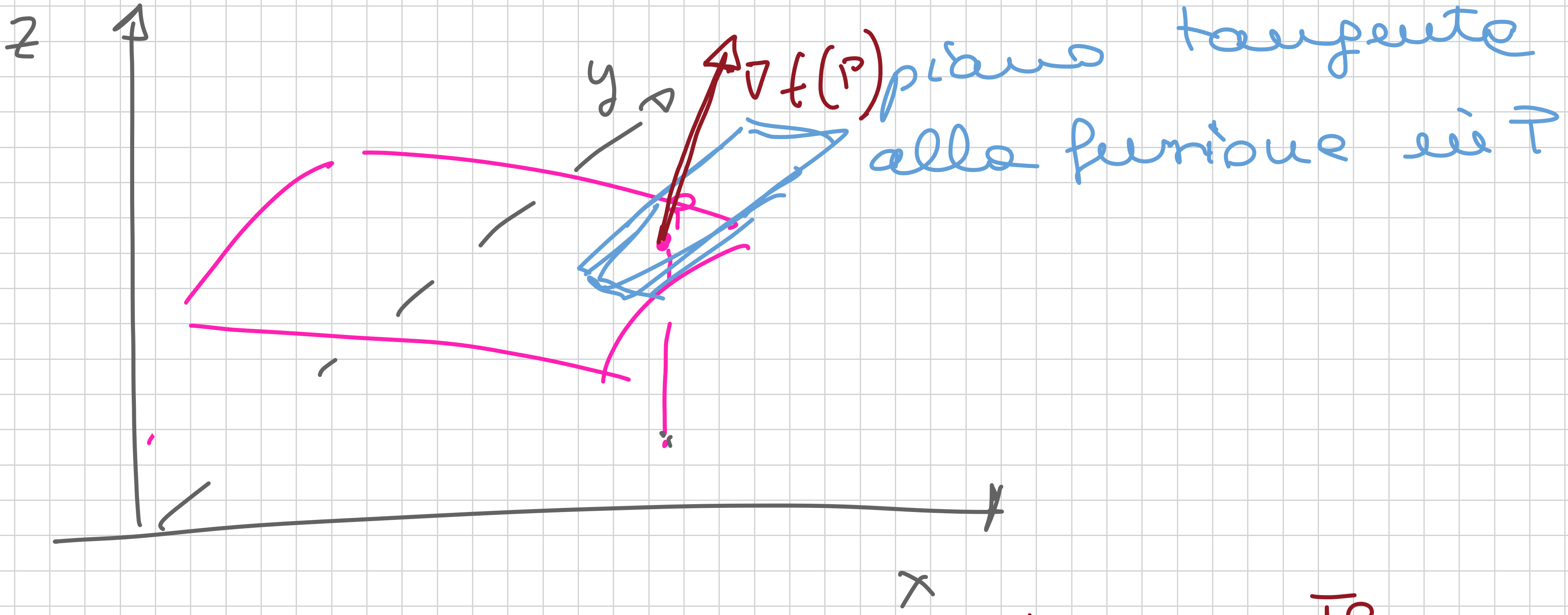
Partiamo con il calcolo delle derivate parziali:

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

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.



$\nabla f(P)$ è al piano tangente

ad f in P con verso uscente

matrice Hessiana

$n \times n$

$$f : R^2 \to R$$

$$\nabla^2 f : R^2 \to R^{2\times 2} \quad \text{matrice } 2\times 2$$

$$\nabla^2 f = \begin{pmatrix} \frac{\partial f^2}{\partial x_1^2} & \frac{\partial f^2}{\partial x_2 x_1} \\ \frac{\partial^2 f}{\partial x_1 x_2} & \frac{\partial^2 f}{\partial x_2^2} \end{pmatrix}$$

$\nabla^2 f$ matrice simmetrica

$f : \mathbb{R}^n \to \mathbb{R}$

$\nabla^2 f : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ matrice $n \times n$

$$\boxed{\nabla^2 f_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}} \qquad i,j = 1 \ldots n$$

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 x_1} & \cdots & \frac{\partial^2 f}{\partial x_n x_1} \\ \frac{\partial^2 f}{\partial x_1 \cancel{x}_2} & & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \frac{\partial^2 f}{\partial x_1 x_n} & & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

$$f(x_1, x_2) = x_1^2 + x_2^2$$

$$\frac{\partial f}{\partial x_1} = 2x_1 \qquad \frac{\partial f}{\partial x_2} = 2x_2$$

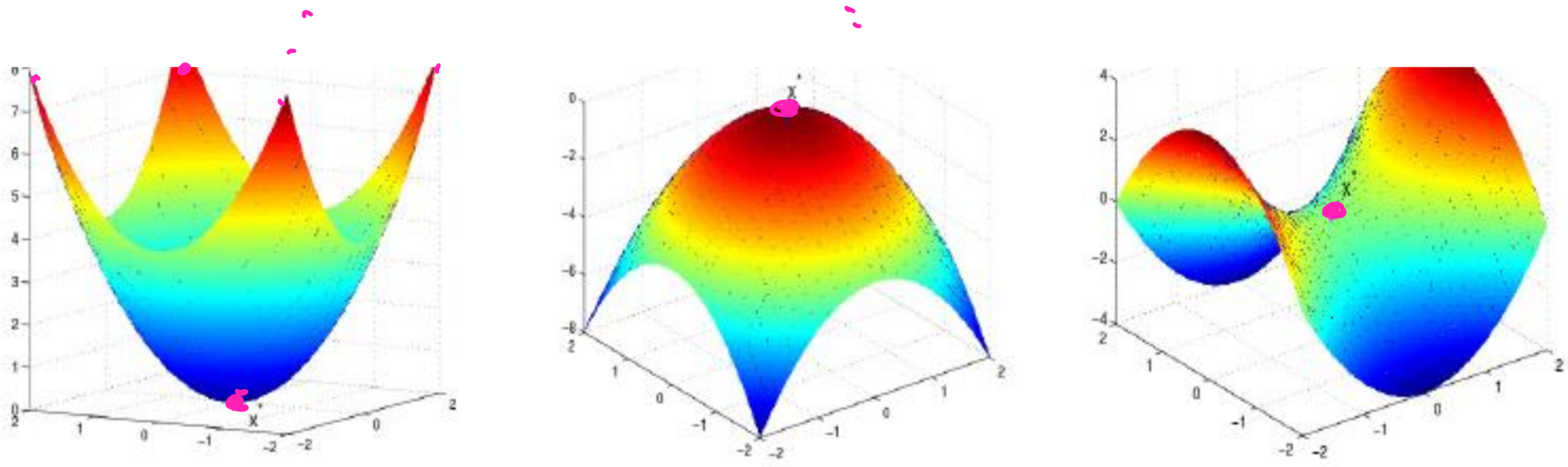
$$\nabla^2 f = \begin{pmatrix} 2 \ \frac{\partial f}{\partial x_1^2} & 0 \ \frac{\partial f}{\partial x_2 \partial x_1} \\ 0 \ \frac{\partial f}{\partial x_1 \partial x_2} & 2 \ \frac{\partial f}{\partial x_2^2} \end{pmatrix}$$

$\nabla^2 f$ è costante solo in casi particolari trovare gli estremi di una funzione

3) Ottimizzazione

STATISTICA NUMERICA, CAP. 6.2.4

$f : \mathbb{R}^2 \to \mathbb{R}$



(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):

$$\boxed{\min_x f(x)} \tag{1}$$

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

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

$f: R^n \to R$

Ottimizzazione :concetti generali

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

$$\boxed{\min_x f(x)} \longleftrightarrow \arg\min_x f(x) \qquad (2)$$

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

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

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

$x \in [x^* - \varepsilon, x^* + \varepsilon]$

$f: \mathbb{R} \to \mathbb{R}$

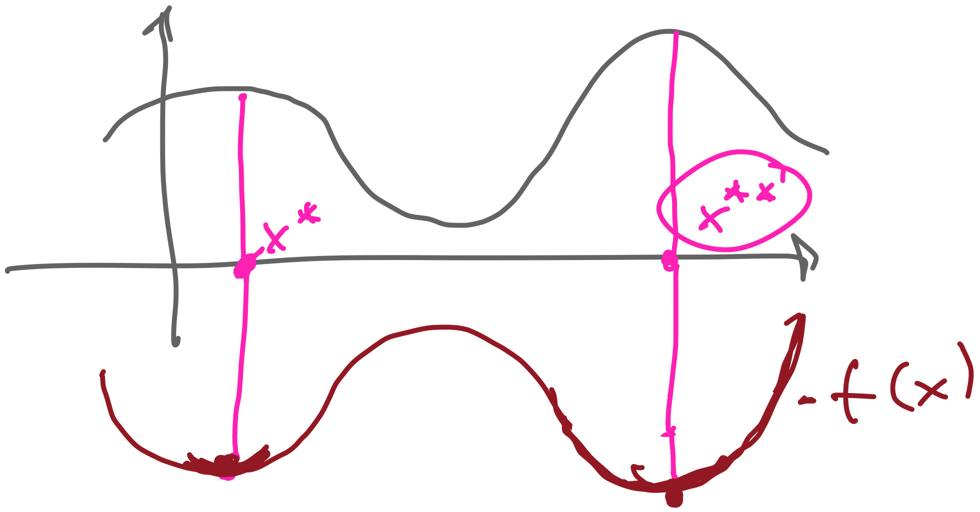
$\arg\max f(x) \to$ $f: \mathbb{R}^n \to \mathbb{R}$

ascissa del punto di massimo

$\Updownarrow$

$\arg\min -f(x)$

ascissa punto di minimo



Ottimizzazione :concetti generali

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

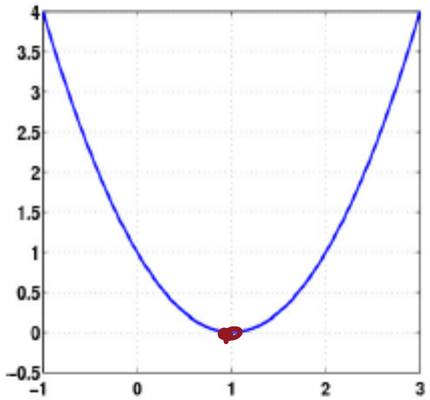
$$\boxed{f(x^*) \leq f(x) \quad \text{per ogni } x \in \mathbb{R}^n.} \tag{4}$$

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

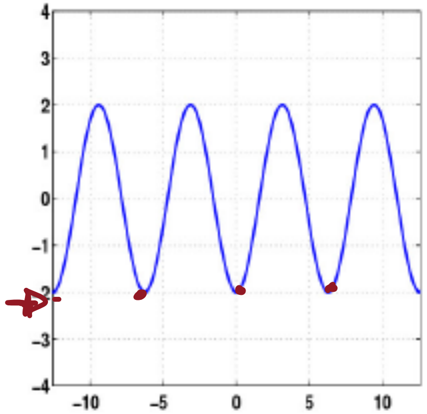
$$f(x^*) < f(x) \quad \text{per ogni } x \in \mathbb{R}^n, \quad x \neq x^*. \tag{5}$$

Ottimizzazione :concetti generali

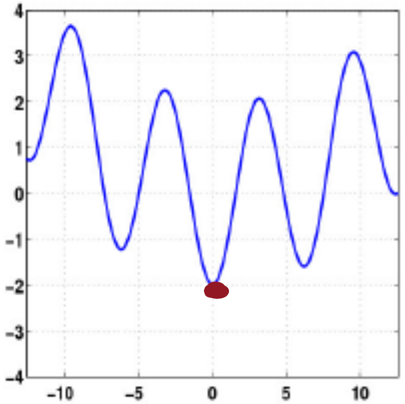
Una funzione $f(x)$ può avere un unico punto di minimo locale (quindi anche globale), oppure può non avere nè minimi locali nè globali, può avere sia minimi locali che globali... (figure 1).



(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

∃ tutte le derivate prime parziali di f

Teorema (Condizioni necessarie del primo ordine)

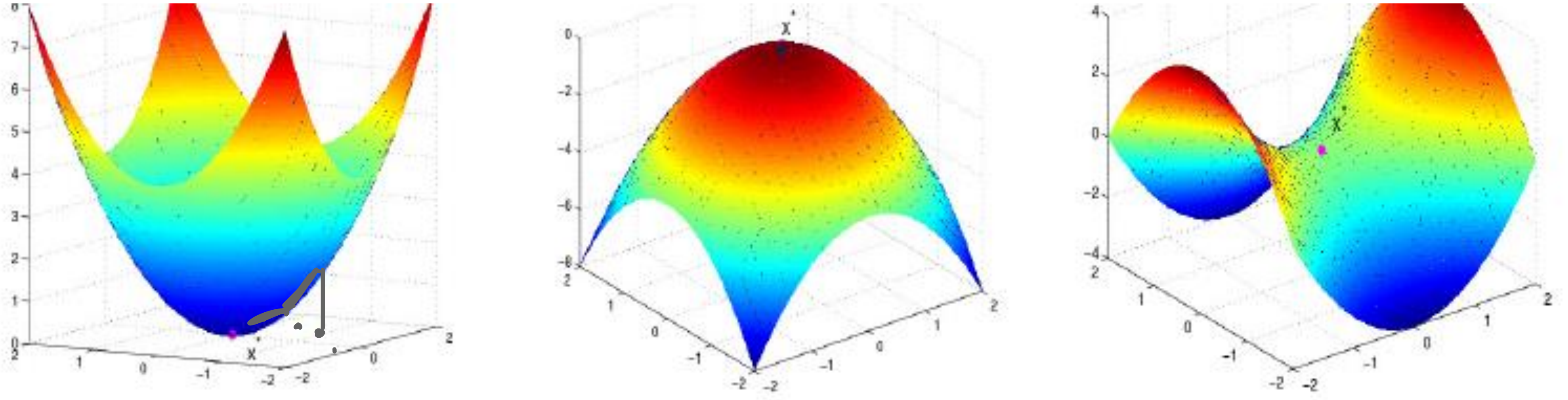
Se x^* è un punto di minimo locale e f è differenziabile con continuità in un intorno aperto di x^*, allora $\nabla f(x^*) = 0$.

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

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

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

$\not\Leftarrow$



(a) Punto di minimo. (b) Punto di massimo. (c) Punto di sella.

Ottimizzazione :concetti generali

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 $x_1, x_2, \ldots$, a partire da un iterato iniziale x_0 assegnato, secondo una legge del tipo:

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

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

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

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.

CRITERI di ARRESTO

$x_0, x_1, x_2 \dots x_k, x_{k+1} \dots | \dots \longrightarrow x^*$

x_0 input $\qquad$ x_{k^*}

$k = 0$

finché criterio = vero

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

$k = k+1$

Metodi di discesa

$x_k \in R^n$

- Si consideri il problema della minimizzazione non vincolata di una funzione $f : \mathbb{R}^n \to \mathbb{R}$ 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

$$x_0, x_1, x_2, \ldots$$

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

definiti dall'iterazione

$$\boxed{x_{k+1} = x_k + \alpha_k p_k} \qquad (2)$$

$\alpha_k \in R$

dove il vettore p_k è una direzione di ricerca e lo scalare α_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 α_k sono scelti in modo da garantire la decrescita di $f(x)$ ad ogni iterazione:

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

Definizione

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

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

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

Metodi di discesa

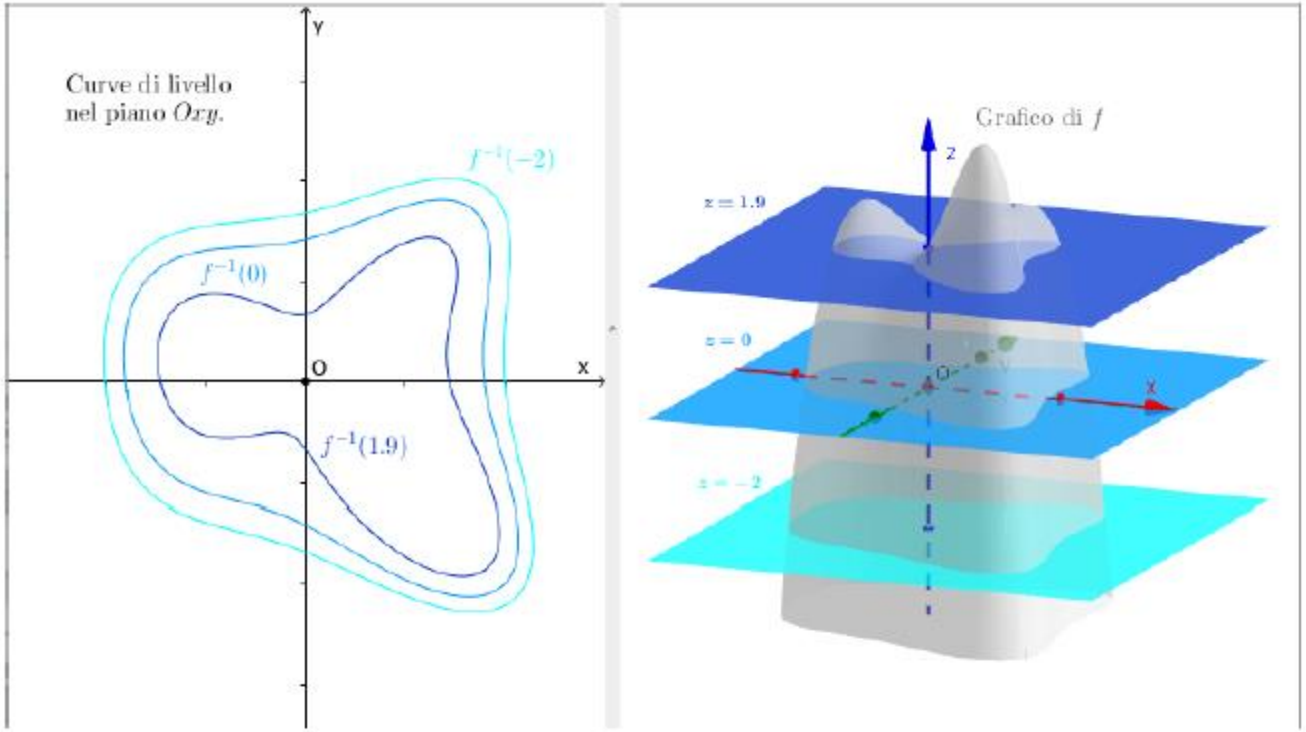
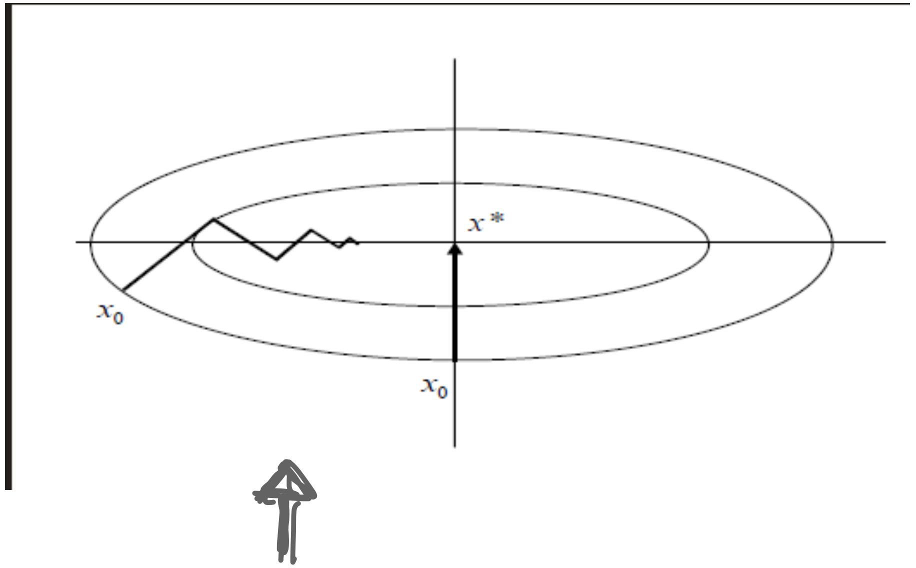


Figura 2.4: Esempio di curve di livello

Metodo del gradiente

$\alpha \sim 10^{-3}, 10^{-4}$

$\alpha < 1$



$\boxed{\begin{array}{l} \text{Algoritmo del metodo del gradiente} \\ \\ \text{Input: f, x\_0, } \alpha \\ \\ \text{k=0} \\ \text{Ripeti fino a convergenza} \\ \qquad x_{k+1} = x_k - \alpha \;\; \nabla f(x_k) \\ \qquad \text{k=k+1} \end{array}}$

$\alpha_k = \alpha \quad \forall k \rightarrow$ α fissato

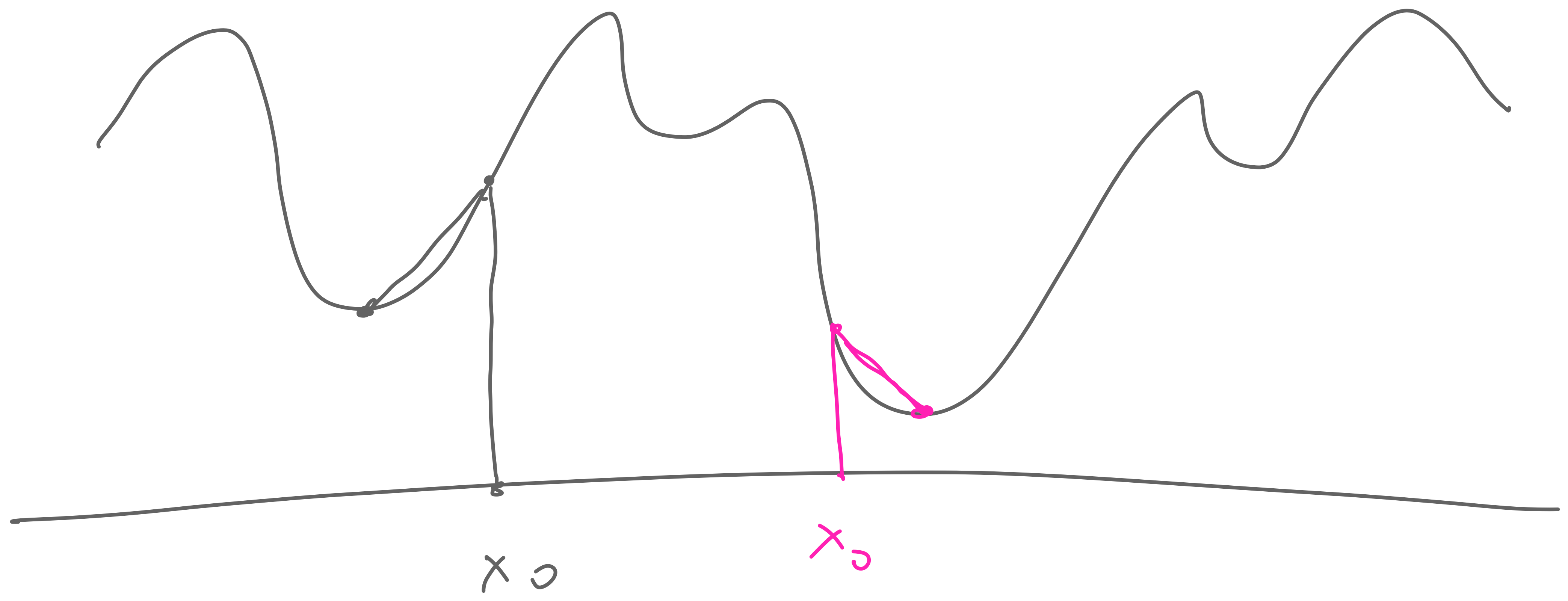
FUNZIONI CONVESSE

Proprietà: punto stazionario che è esiste un solo minimo locale che quindi è anche minimo globale.

FUNZIONI NON CONVESSE

esistono più punti stazionari che possono essere minimi locali, minimi globali o punti di sella.

Il metodo del gradiente converge a un punto di minimo locale, o minimo globale o di sella.


x0
x0

Metodi di discesa

- 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

$$\boxed{\|\nabla f(x_k)\| \leq \epsilon, \quad \epsilon > 0} \qquad (1)$$

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