Machine Learning

Anno Accademico 2021 - 2022

Richiami di Matematica

Sommario

- Richiami sulle Funzioni Convesse
- Funzioni di più Variabili (Derivate Parziali)
- Gradiente di una Funzione
- Algoritmo di Gradient Descent
- Cenni di Calcolo delle Probabilità

Richiami sulle Funzioni Convesse

Insiemi Convessi

Un insieme C in uno spazio vettoriale è *convesso* se, comunque si scelgano due punti $\mathbf{v}$ e $\mathbf{w}$ appartenenti a C, il segmento che unisce i due punti appartiene a C.

Più formalmente:

Un insieme C in uno spazio vettoriale è *convesso* se, $\forall\ \mathbf{v}, \mathbf{w} \in C$, e $\forall\ \lambda \in [0, 1]$, si ha:

$$\lambda\mathbf{v} + (1 - \lambda)\mathbf{w} \in C$$

Insiemi Convessi

L'espressione:

$$\lambda\mathbf{v} + (1 - \lambda)\mathbf{w}$$

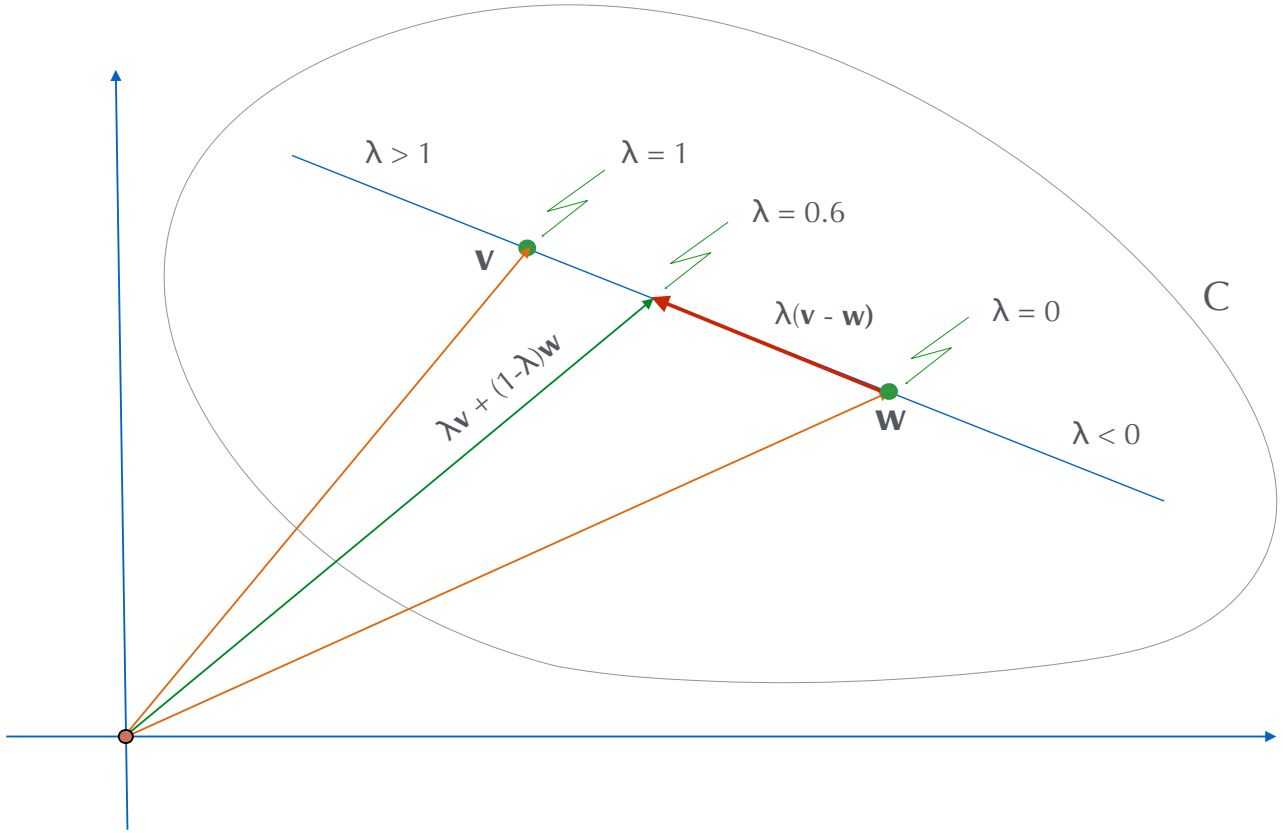
corrisponde dunque ai punti appartenenti al segmento che unisce i due punti **v** e **w**, al variare di $\lambda \in [0, 1]$.

Vediamolo nel caso a due dimensioni:

Insiemi Convessi

[caso a due dimensioni]

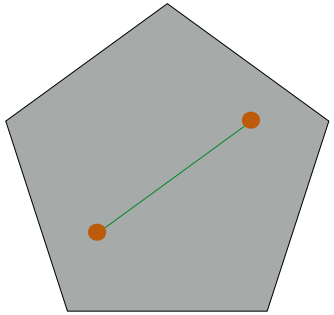
$\lambda\mathbf{v} + (1-\lambda)\mathbf{w} = \mathbf{w} + \lambda(\mathbf{v}-\mathbf{w})$



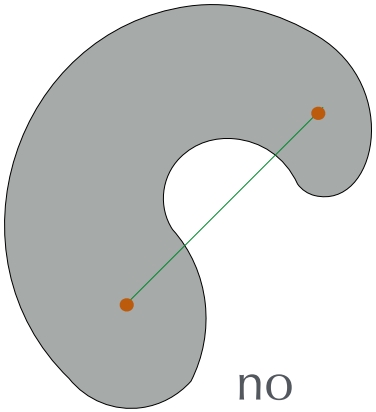
Insiemi Convessi

[caso a due dimensioni]

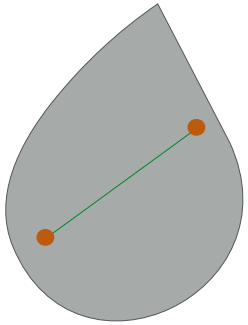
Vediamo ora alcuni esempi di insiemi convessi e non convessi nel caso di spazio a due dimensioni:



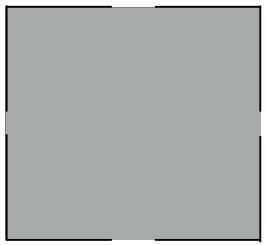
si



no



si

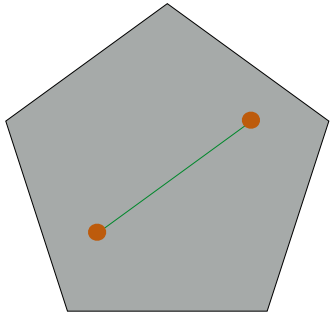


?

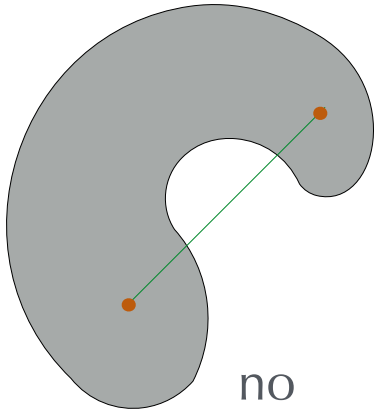
Insiemi Convessi

[caso a due dimensioni]

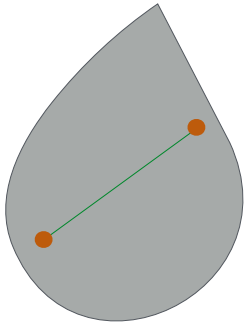
Vediamo ora alcuni esempi di insiemi convessi e non convessi nel caso di spazio a due dimensioni:



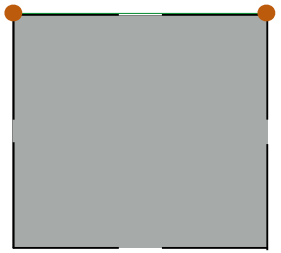
si



no



si



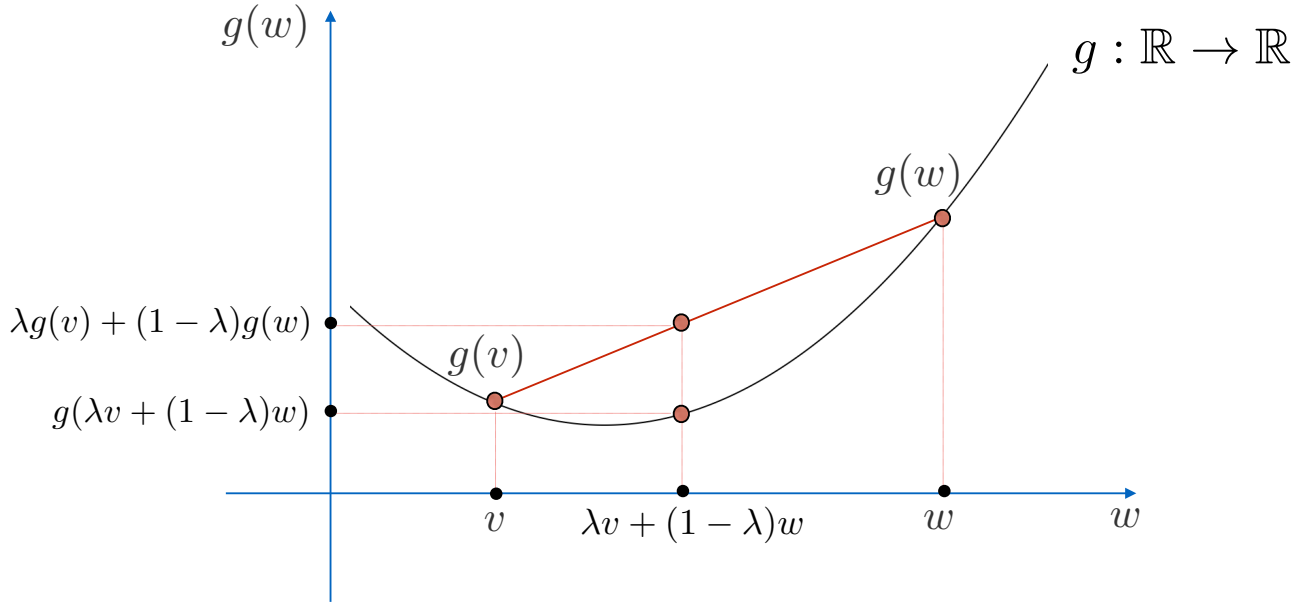
no

Funzioni Convesse

Sia C un insieme convesso. Una funzione $g : C \to \mathbb{R}$ si dice convessa se, per ogni **v** e **w** appartenenti al suo dominio di definizione, vale la seguente proprietà:

$$g(\lambda \mathbf{v} + (1 - \lambda)\mathbf{w}) \leq \lambda g(\mathbf{v}) + (1 - \lambda) g(\mathbf{w})$$

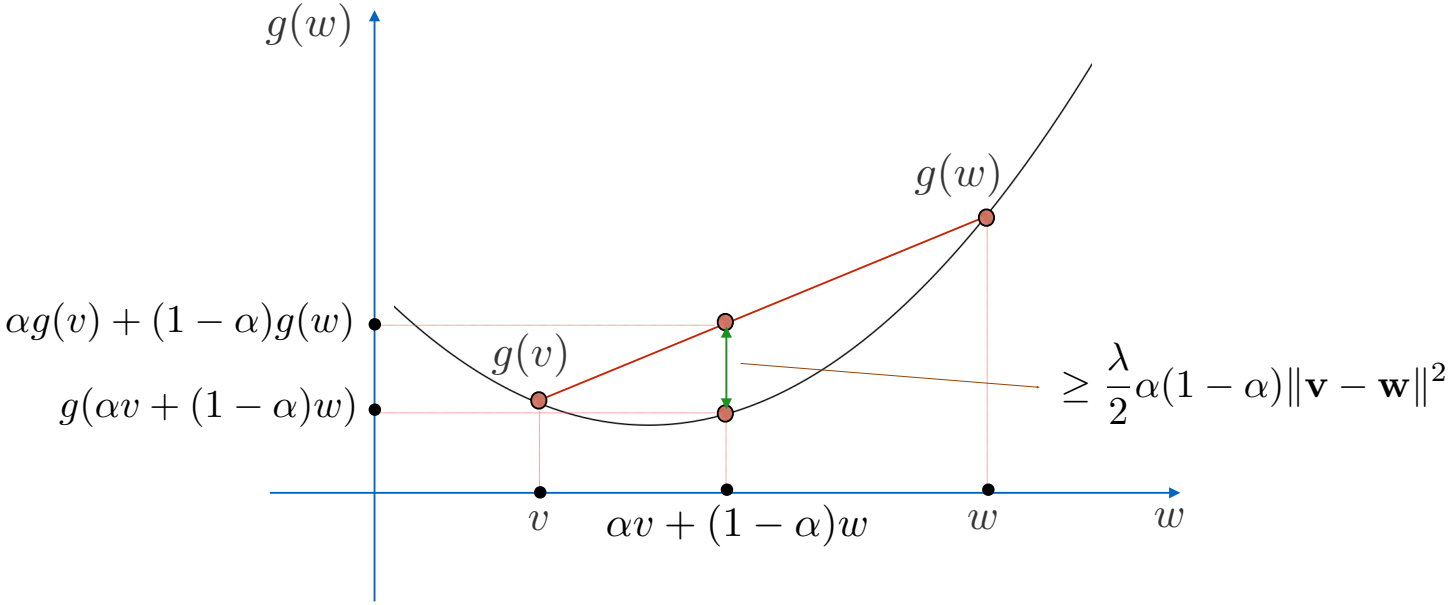
$$\text{con} \quad \lambda \in [0, 1]$$



Definizione di Funzione Strongly Convex

- Una funzione g è detta λ-*strongly convex* se, per ogni $\mathbf{w}$, $\mathbf{v}$ e $\alpha \in (0, 1)$, si ha:

$$g(\alpha\mathbf{v} + (1-\alpha)\mathbf{w}) \leq \alpha g(\mathbf{v}) + (1-\alpha)g(\mathbf{w}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{v}-\mathbf{w}\|^2$$

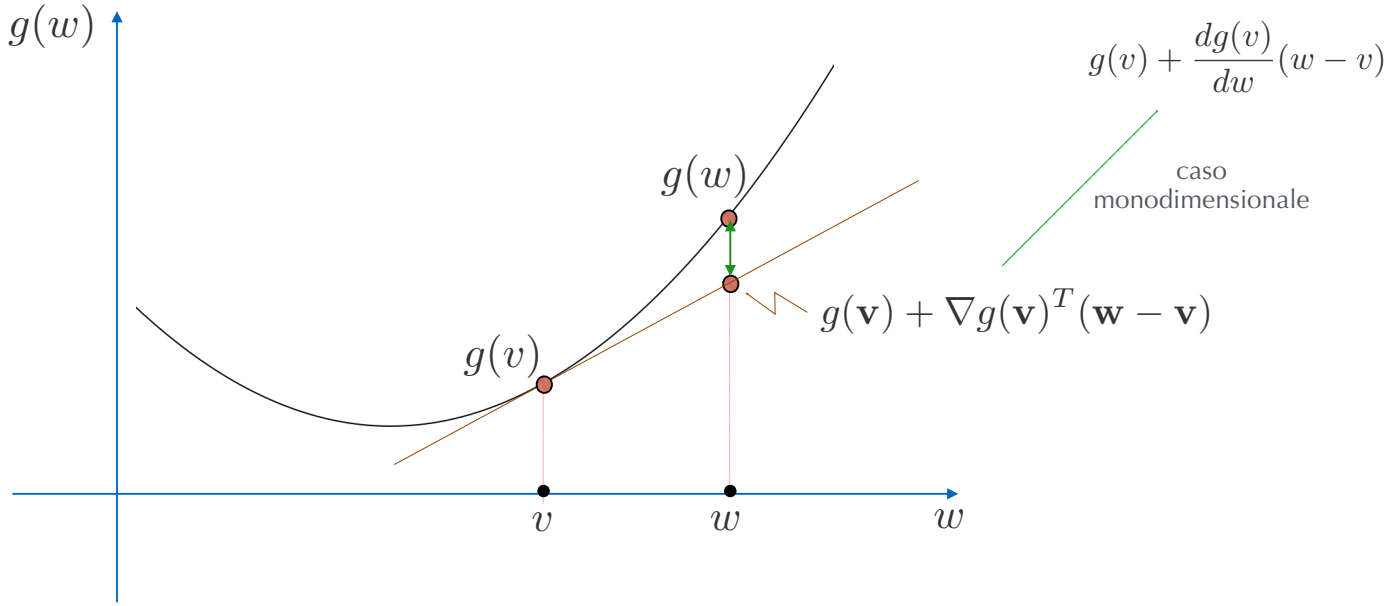


- Ovviamente, ogni funzione convessa è 0-*strongly convex*.

Funzioni Convesse

Per una funzione convessa g(**w**) differenziabile, il "piano" tangente giace sempre al di sotto del grafico della funzione:

$$g(\mathbf{w}) \geq g(\mathbf{v}) + \nabla g(\mathbf{v})^T(\mathbf{w} - \mathbf{v})$$



Funzioni di più Variabili

Derivate Parziali di Funzioni di più Variabili

- Consideriamo una funzione $g(w_0, w_1)$ di 2 variabili definita in un campo A.
- Sia $\overline{P} \equiv (\overline{w}_0, \overline{w}_1) \in A$. Esiste allora un intorno circolare di centro $\overline{P}$ e opportuno raggio σ, contenuto in A. Ne segue che, se:

$$0 < |\Delta w_0| < \sigma$$

si ha $(\overline{w}_0 + \Delta w_0, \overline{w}_1) \in A$ e si può considerare il seguente rapporto incrementale:

$$\frac{g(\overline{w}_0 + \Delta w_0, \overline{w}_1) - g(\overline{w}_0, \overline{w}_1)}{\Delta w_0}$$

che si chiama *rapporto incrementale parziale rispetto a* w_0 della g , perché in esso consideriamo incrementata soltanto la w_0, mantenendo inalterata la w_1 .

Derivate Parziali di Funzioni di più Variabili

- Se esiste determinato e finito il seguente limite:

$$\lim_{\Delta w_0 \to 0} \frac{g(\overline{w}_0 + \Delta w_0, \overline{w}_1) - g(\overline{w}_0, \overline{w}_1)}{\Delta w_0}$$

la funzione g si dice *parzialmente derivabile rispetto a* w_0 nel punto $(\overline{w}_0, \overline{w}_1)$.

Derivate Parziali di Funzioni di più Variabili

- Supponiamo ora che la funzione g sia parzialmente derivabile rispetto a w_0 in ogni punto del campo A.
- Per ogni punto di A resta ben determinato il corrispondente valore della derivata parziale rispetto a w_0.
- Nasce così in A una nuova funzione di due variabili w_0, w_1 che si chiama *derivata parziale rispetto a* w_0 *della funzione g* e si denota ad esempio come segue:

$$\frac{\partial g}{\partial w_0}$$

Derivate Parziali di Funzioni di più Variabili

- Analogamente si definisce la derivata parziale rispetto a w_1, nel punto $\overline{P}$, come il limite

$$\lim_{\Delta w_1 \to 0} \frac{g(\overline{w}_0, \overline{w}_1 + \Delta w_1) - g(\overline{w}_0, \overline{w}_1)}{\Delta w_1}$$

supposto determinato e finito.

- E se avviene che tale derivata esista in ogni punto $(w_0, w_1) \in A$, resta ivi definita una nuova funzione delle variabili w_0, w_1 che si chiama la derivata parziale rispetto a w_1 della $g(w_0, w_1)$ e si indica ad esempio come segue:

$$\frac{\partial g}{\partial w_1}$$

Derivate Parziali di Funzioni di più Variabili

- Osserviamo che, mentre per le funzioni di una variabile la derivabilità in un punto implica la continuità in tale punto, non sussiste il fatto analogo per le funzioni di due variabili.

- Possono cioè in un punto esistere le due derivate parziali senza che la funzione g sia continua in esso.

- Tutte le considerazioni fatte fino ad ora sulle funzioni di due variabili si estendono immediatamente al caso delle funzioni di più di due variabili:

$$g(w_0, w_1, \ldots, w_n) = g(\mathbf{w})$$

Gradiente

Gradiente di una Funzione

- Il gradiente di una funzione è una diretta generalizzazione della nozione di derivata per una funzione a più variabili.
- Data la seguente funzione:

$$g(w_0, w_1, \ldots, w_n) = g(\mathbf{w})$$

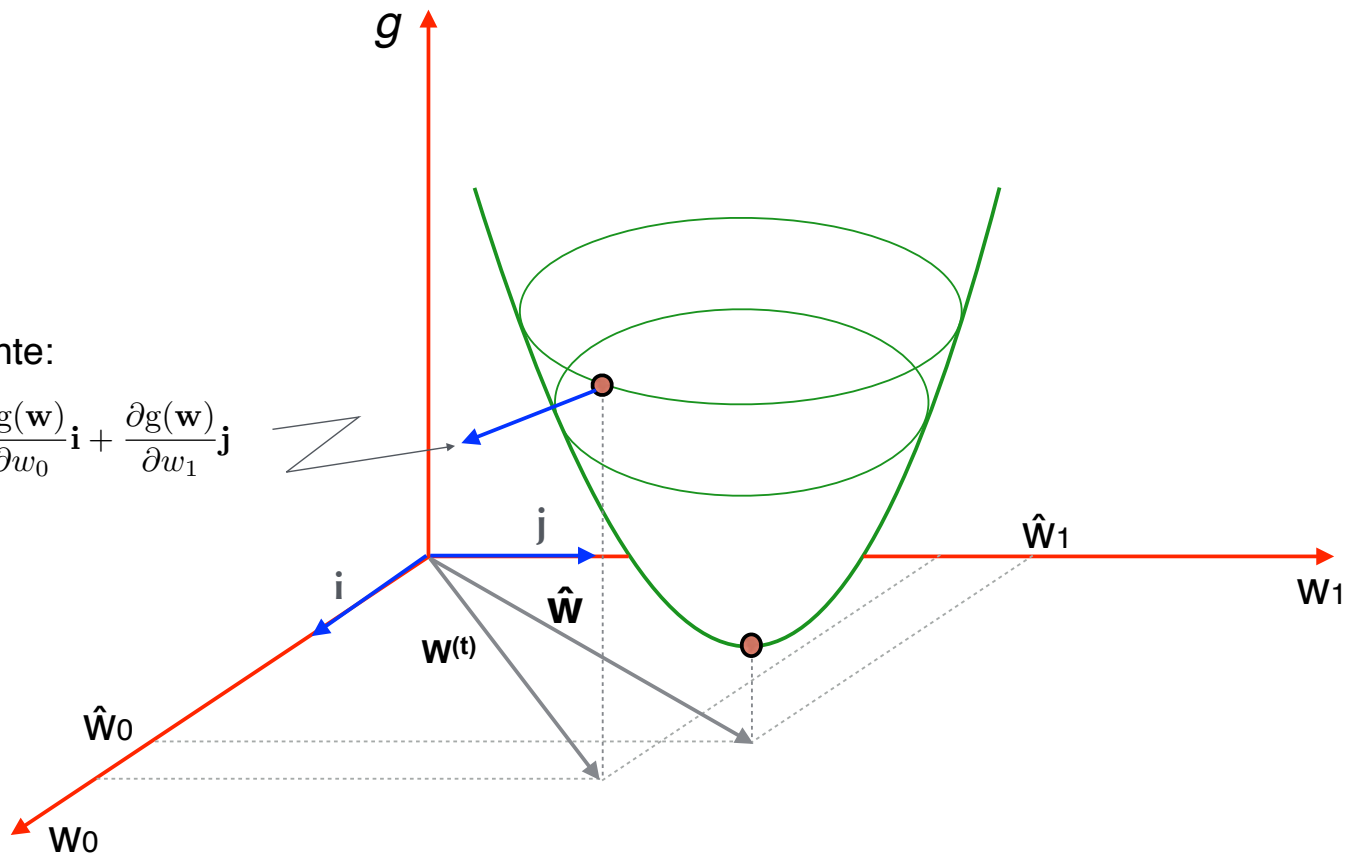
definiamo gradiente di g il vettore le cui componenti sono le derivate parziali della funzione:

$$\nabla \mathrm{g}(\mathbf{w}) = \begin{bmatrix} \frac{\partial \mathrm{g}(\mathbf{w})}{\partial w_0} \\ \\ \frac{\partial \mathrm{g}(\mathbf{w})}{\partial w_1} \\ \\ \cdots \\ \\ \frac{\partial \mathrm{g}(\mathbf{w})}{\partial w_n} \end{bmatrix}$$

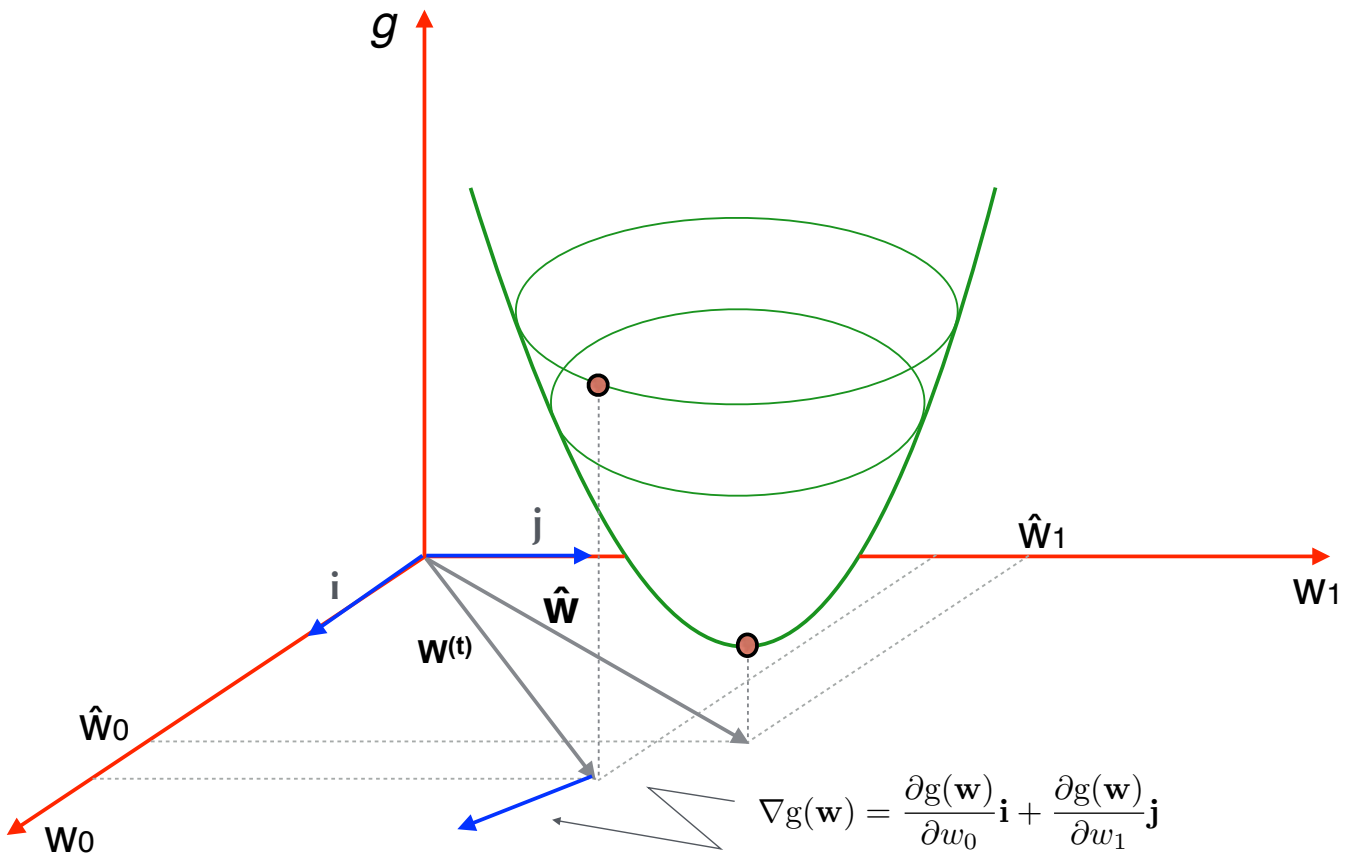
Gradiente di una Funzione

gradiente:

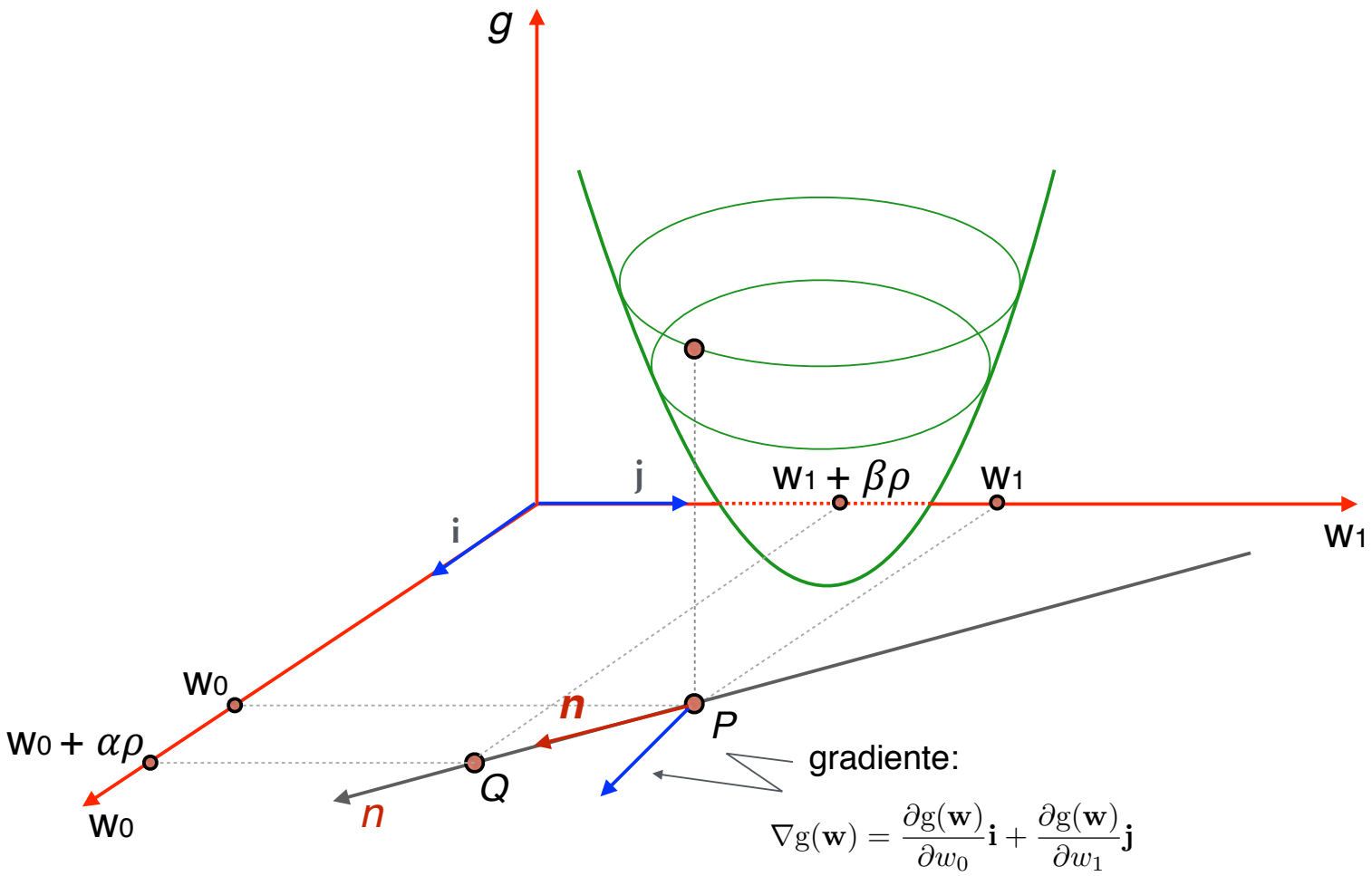
$$\nabla g(\mathbf{w}) = \frac{\partial g(\mathbf{w})}{\partial w_0}\mathbf{i} + \frac{\partial g(\mathbf{w})}{\partial w_1}\mathbf{j}$$



Gradiente di una Funzione



Derivata Direzionale



Derivata Direzionale

- Si può dimostrare che la derivata direzionale secondo n è:

$$\frac{\partial g}{\partial n} = \alpha \cdot \frac{\partial g}{\partial w_0} + \beta \cdot \frac{\partial g}{\partial w_1}$$

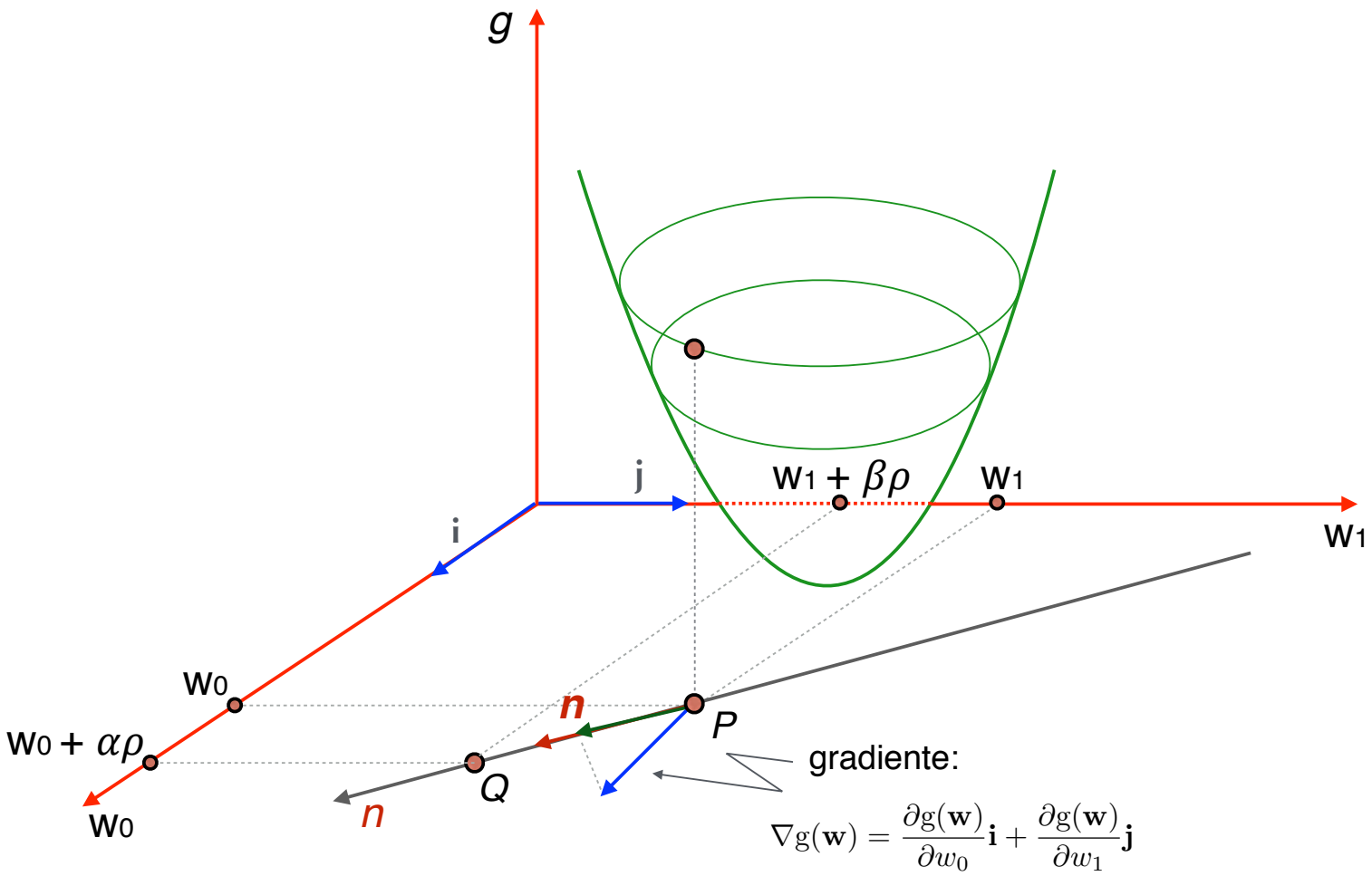
Detto $\boldsymbol{n}$ il versore della direzione *n*, la $\frac{\partial g}{\partial n}$ è il prodotto scalare dei due vettori:

$$\nabla g \cdot \boldsymbol{n}$$

cioè è la componente del gradiente sulla retta orientata *n*. Questo significa che la derivata direzionale della funzione *g* è massima secondo la direzione e verso del vettore gradiente.

- Si può avere una visione globale di tutte queste possibili derivate, collegando al punto *P* il gradiente della funzione g in tale punto.

Derivata Direzionale



Algoritmo Gradient Descent

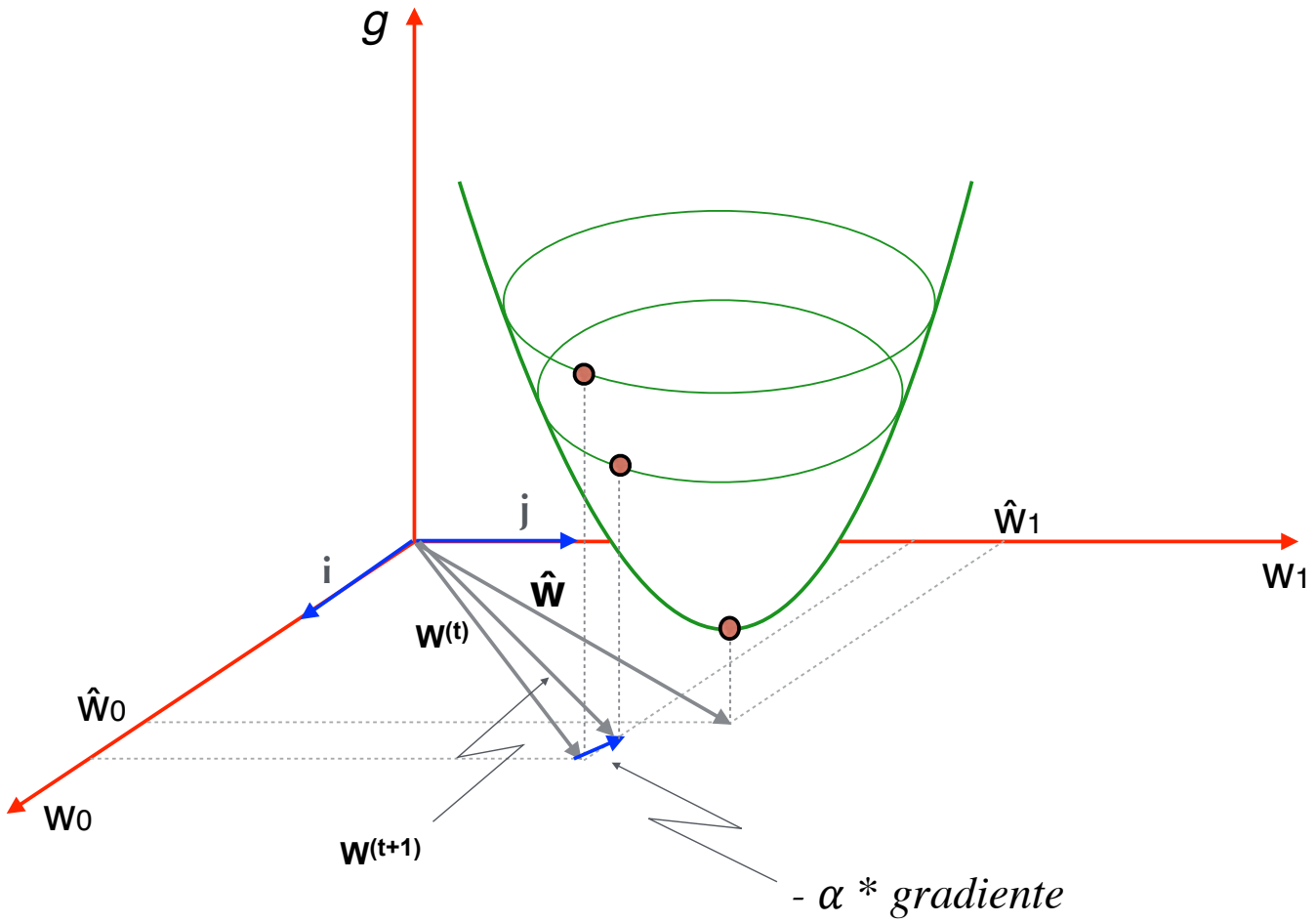
- La proprietà citata in precedenza del vettore gradiente, ossia il fatto che il gradiente fornisce direzione della pendenza più ripida, è alla base di algoritmi di Ricerca Locale che operano in spazi continui.

- Tali algoritmi si dividono in due classi principali:

  - Algoritmi a Salita più Ripida (Hill-Climbing)

  - Algoritmi a Discesa del Gradiente (Gradient Descent)

Algoritmo Gradient Descent



Algoritmo Gradient Descent

$\mathbf{w}^{(1)} = 0$ (oppure lo inizializziamo in modo casuale)

$t = 1$

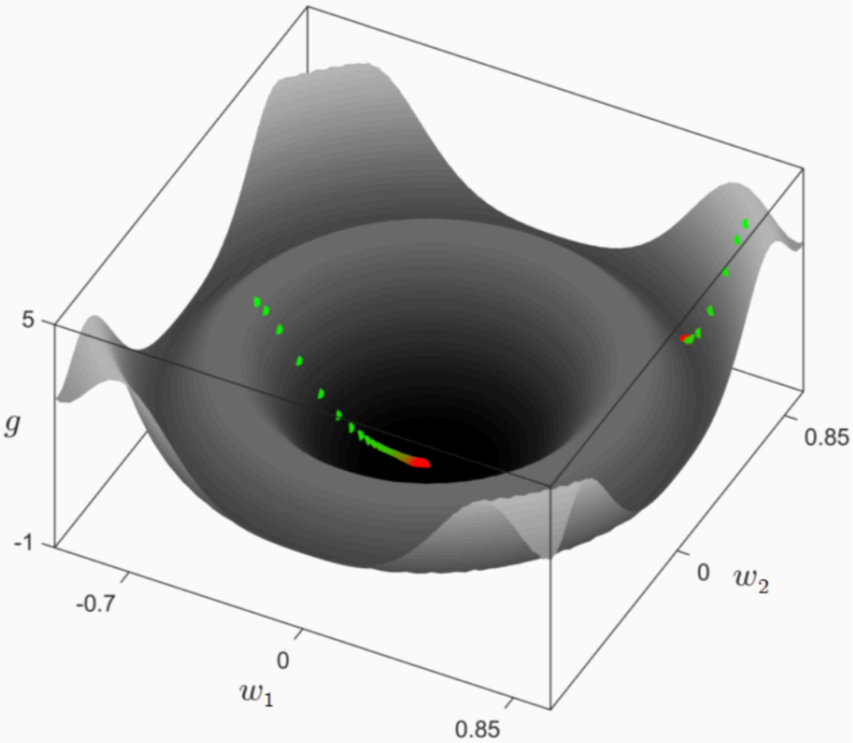
while $\|\nabla \mathrm{g}(\mathbf{w}^{(t)})\|_2 > \epsilon$

$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \alpha * \nabla \mathrm{g}(\mathbf{w}^{(t)})$

$t \leftarrow t + 1$

Algoritmo Gradient Descent

- Funzione non convessa di due variabili:



Richiami di Probabilità

Variabili Aleatorie

- Le quantità di interesse che sono determinate dal risultato di un esperimento casuale sono dette *variabili aleatorie*.
- Poiché il valore di una variabile aleatoria è determinato dall'esito di un esperimento, possiamo assegnare delle probabilità ai suoi valori possibili.
- Esempi di v.a.: risultato del lancio di un dado, risultato del lancio di una moneta, ecc.

Valore Atteso

- Il concetto di Valore Atteso è uno dei più importanti concetti in tutta la teoria della probabilità.
- Sia X una variabile aleatoria discreta che può assumere i valori x_1, x_2, …, x_N. Il Valore Atteso di X è il numero:

$$E[X] \triangleq \sum_{i=1}^{N} [x_i \cdot P(X = x_i)]$$

Valore Atteso

- Si tratta della media pesata dei valori possibili di X, usando come pesi le probabilità che tali valori vengano assunti da X.
- Per questo E[X] è anche detto *media* di X (termine che però è sconsigliabile), oppure *aspettazione* (*expectation*).

Valore Atteso

ESEMPIO: lancio di un dado

- Sia X il punteggio che si ottiene lanciando un dado non truccato. Quanto vale E[X]?

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2} = 3.5$$

Valore Atteso

ESEMPIO: lancio di un dado

- Si noti che in questo esempio il valore atteso di X non è uno dei possibili valori che X può assumere.
- Perciò, anche se E[X] è chiamato *valore atteso* di X, ciò non vuole affatto dire che noi ci attendiamo di vedere questo valore, ma piuttosto che ci aspettiamo che sia il limite a cui tende il punteggio medio del dado su un numero crescente di ripetizioni.

Valore Atteso

ESEMPIO: Indicator Function

- Se I[A] è la funzione indicatrice di un evento A, ossia se:

$$I[A] \triangleq \begin{cases} 1 & \text{se A si verifica} \\ 0 & \text{se A non si verifica} \end{cases}$$

allora:

$$E[I] = 1 \cdot P(I = 1) + 0 \cdot P(I = 0) = P(I = 1) = P(A)$$

- Quindi il valore atteso della indicator function di un evento è la probabilità di quest'ultimo.

Valore Atteso

Proprietà di E

- Si riportano qui di seguito alcune proprietà della funzione E (a e b sono variabili aleatorie):

$$
\begin{aligned}
E[a+b] &= E[a]+E[b] \\
E[k\cdot a] &= k\cdot E[a] \text{ (k costante)} \\
E[a\cdot b] &= E[a]\cdot E[b] \text{ (a e b indipendenti)}
\end{aligned}
$$

Varianza

- Sia X una variabile aleatoria con media μ. La varianza di X è la quantità:

$$\mathrm{Var}(X) \triangleq E[(X-\mu)^2]$$

Varianza

- Esiste una formula alternativa per la varianza, che si ricava in questo modo:

$$\begin{aligned} \mathrm{Var}(X) &\triangleq E[(X-\mu)^2] = \\ &= E[X^2 - 2\mu X + \mu^2] = \\ &= E[X^2] - 2\mu \cdot E[X] + \mu^2 = \\ &= E[X^2] - \mu^2 \end{aligned}$$

ossia:

$$\mathrm{Var}(X) = E[X^2] - E[X]^2$$