Metodi di ottimizzazione matematica

Vincenzo Bonifaci

IN550 – Machine Learning

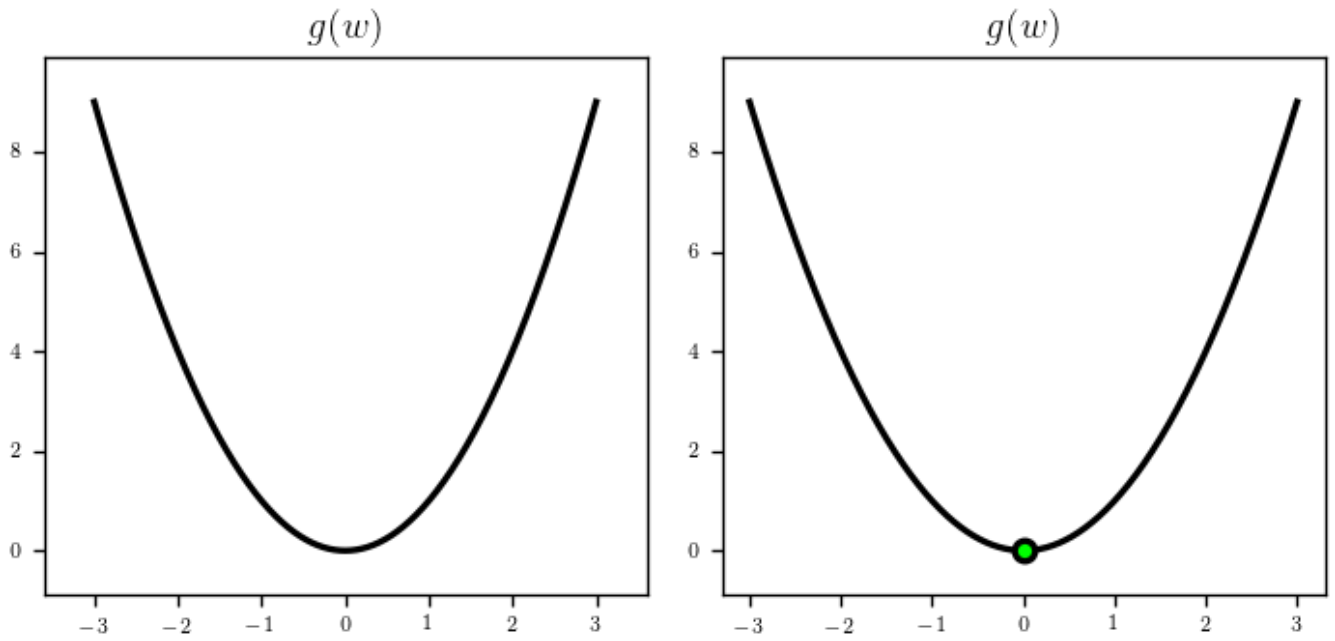
Ottimizzazione matematica

Problema di minimizzazione: $\text{minimize}_{w \in \mathbb{R}^N} \; g(w)$

Input: Una funzione $g : \mathbb{R}^N \to \mathbb{R}$
Output: $w^* \in \mathbb{R}^N$ tale che $g(w^*) \leq g(w)$ per ogni $w \in \mathbb{R}^N$

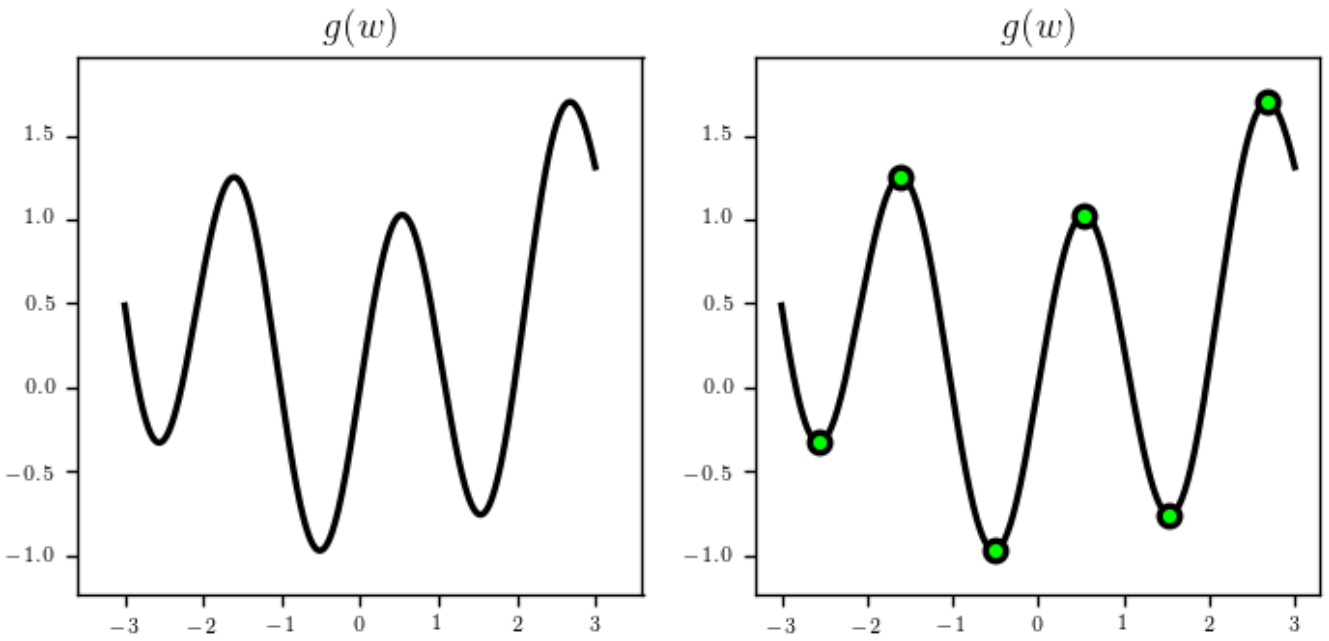
w^* è un *minimo globale* della funzione g

Minimi globali



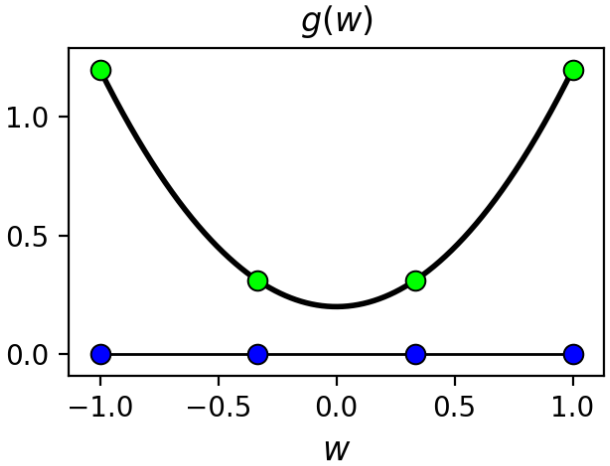
$w^* \in \mathbb{R}^N$ tale che $g(w^*) \leq g(w)$ per ogni w in $\mathbb{R}^N$

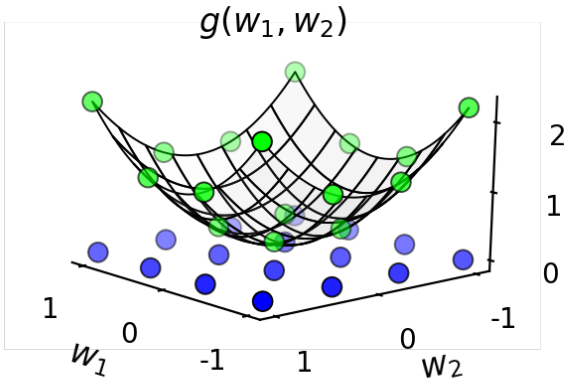
Minimi locali

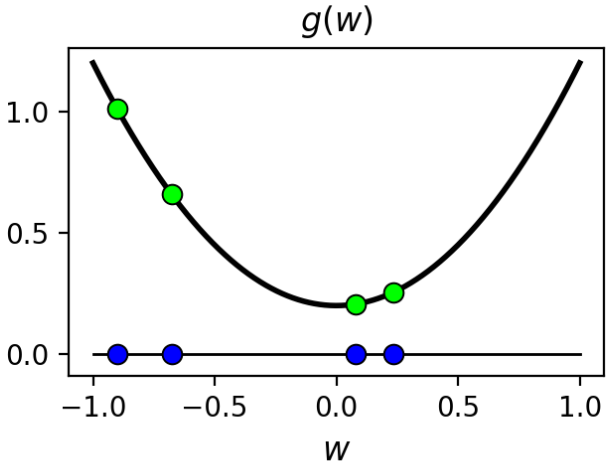


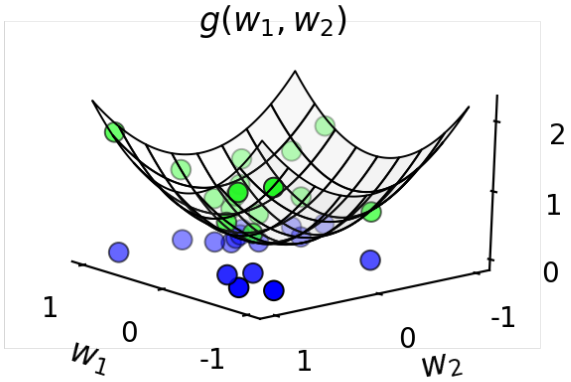
$w^* \in \mathbb{R}^N$ tale che $g(w^*) \leq g(w)$ per ogni w in un intorno di w^*

Due semplici metodi di approssimazione

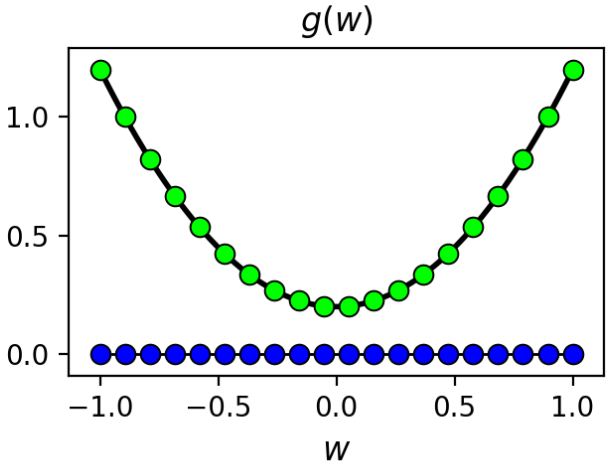


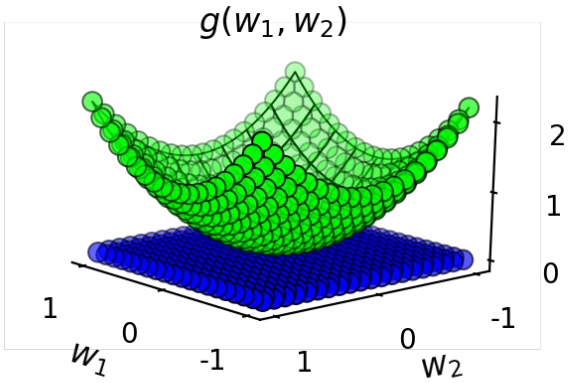


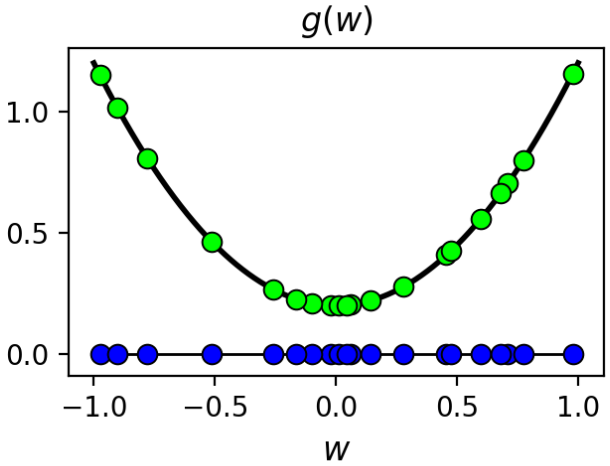


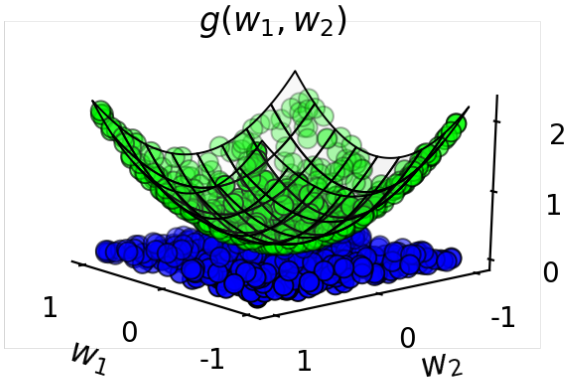


Due semplici metodi di approssimazione

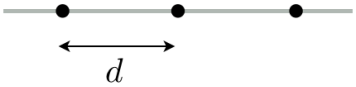


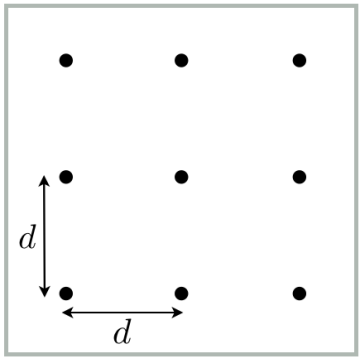


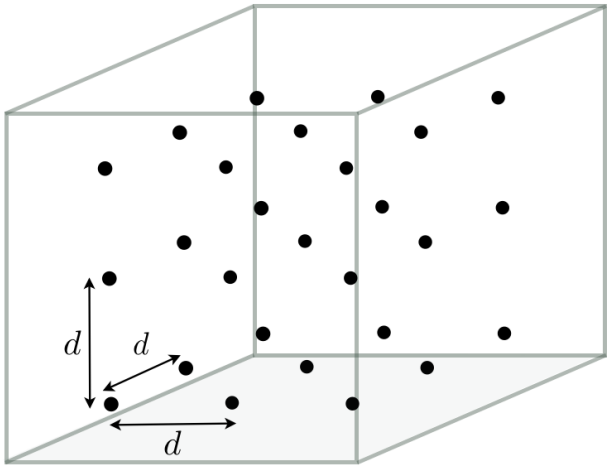




La maledizione della multidimensionalità

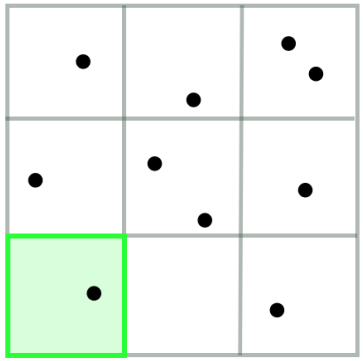




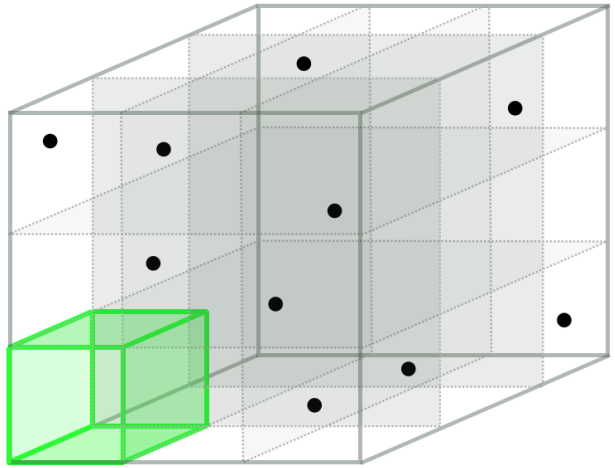


La maledizione della multidimensionalità

$3/10$

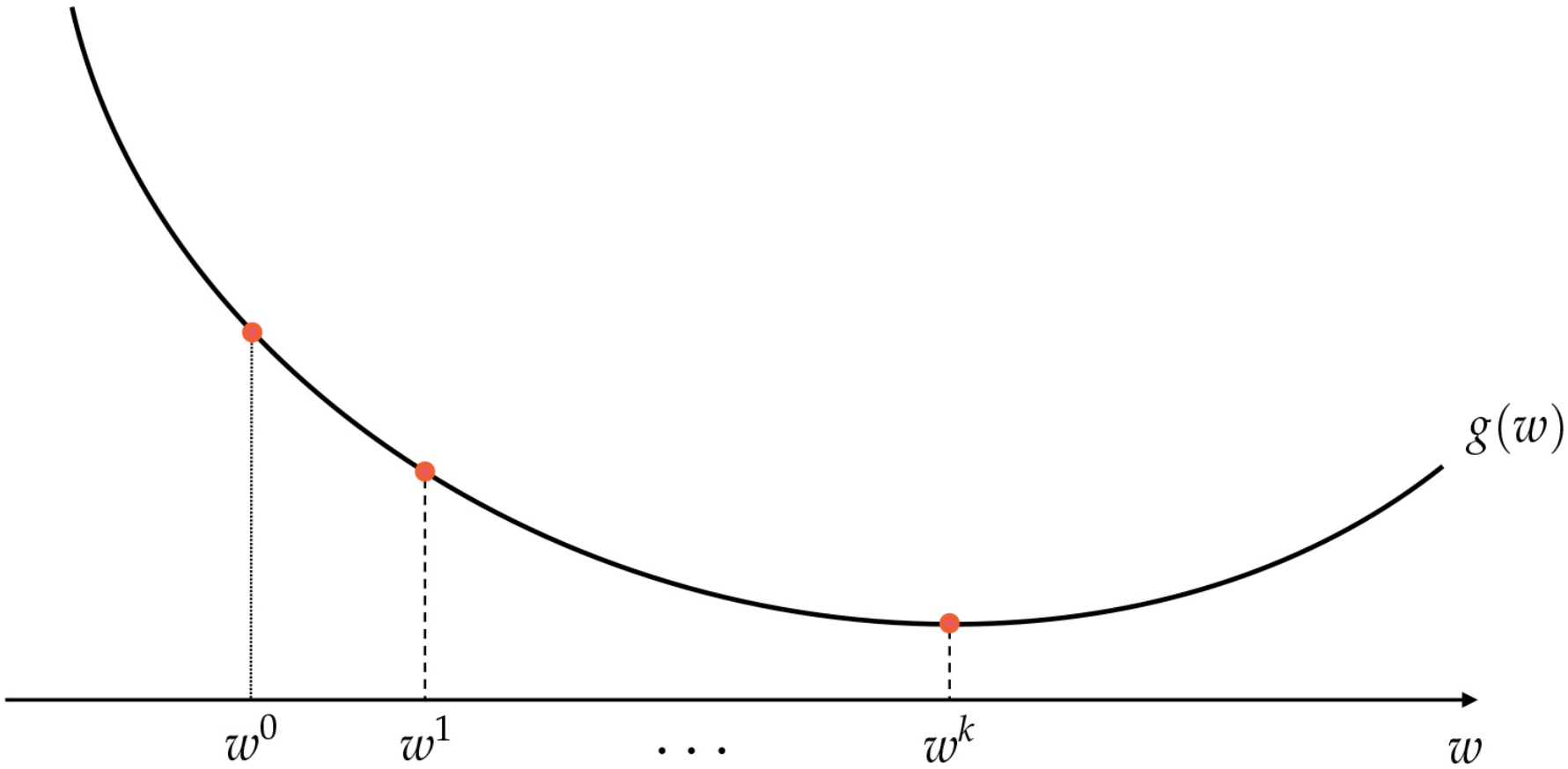


$1/10$

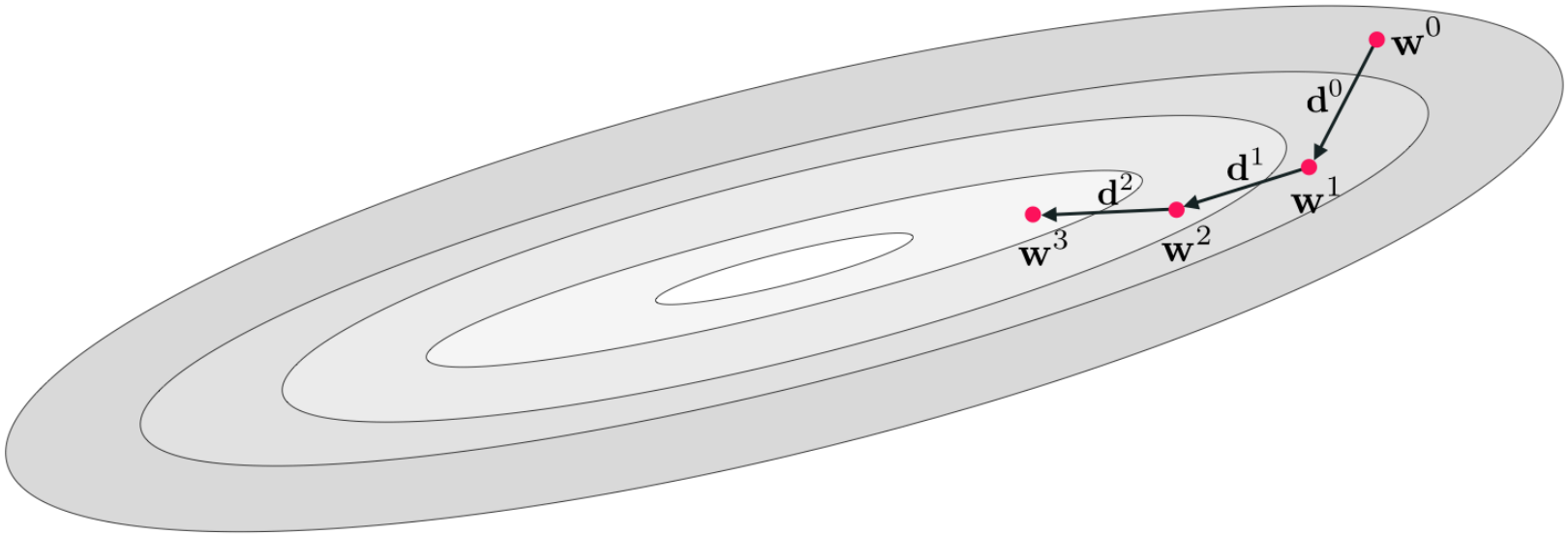


$0/10$

Metodi di ricerca locale



Metodi di ricerca locale

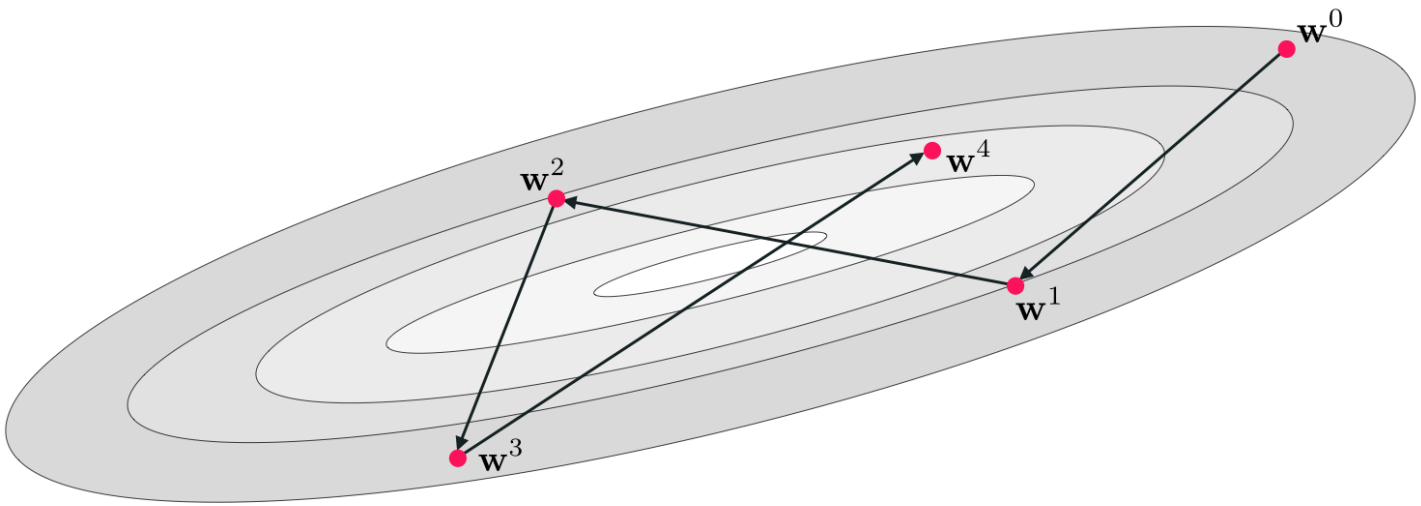


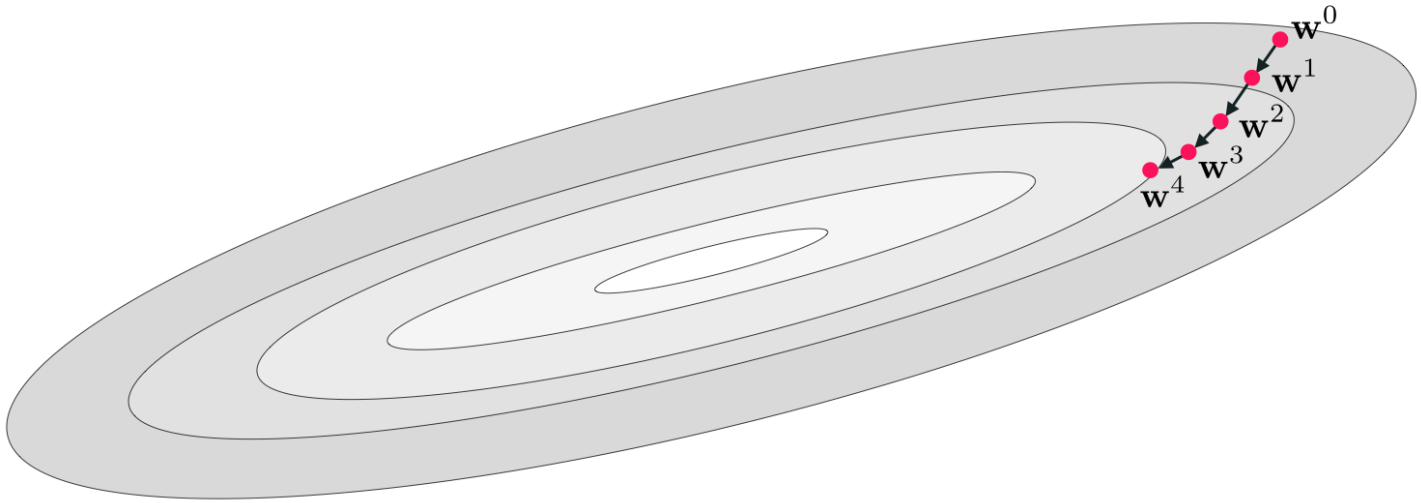
$$w^{(t+1)} = w^{(t)} + d^{(t)}$$

Una condizione desiderabile (non sempre soddisfatta) è che

$$g(w^{(0)}) > g(w^{(1)}) > g(w^{(2)}) > \ldots > g(w^{(t)}) > \ldots$$

Lunghezza del passo





Direzione di discesa e passo

Per controllare la lunghezza del passo possiamo porre, più in generale,

$$w^{(t+1)} = w^{(t)} + \eta\, d^{(t)}$$

- $d^{(t)}$ è la *direzione di discesa* all'iterazione t
- $\eta > 0$ è il *parametro del passo* (*tasso di apprendimento* nel ML)

Poiché

$$\left\| w^{(t+1)} - w^{(t)} \right\| = \left\| \eta d^{(t)} \right\| = \eta \left\| d^{(t)} \right\|$$

la lunghezza del passo è direttamente proporzionale a η

Esempio di algoritmo di discesa: Randomized Local Search

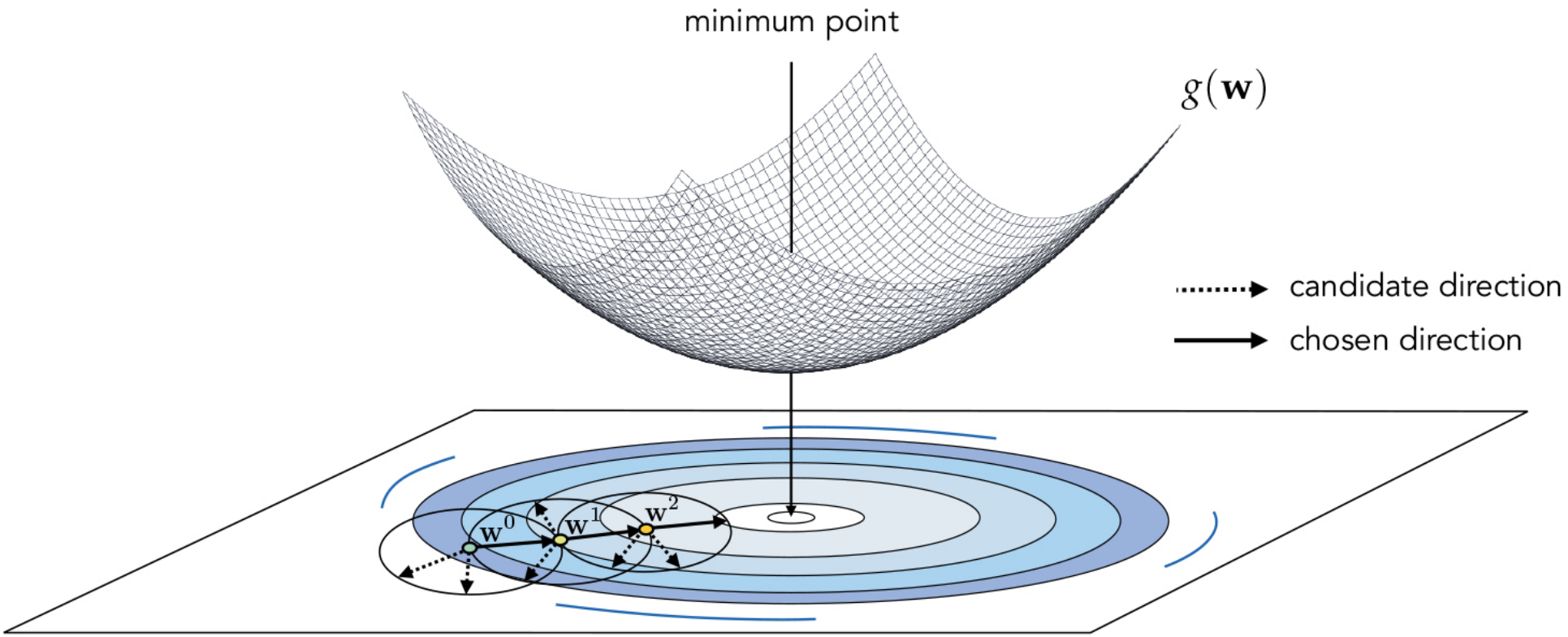
Randomized Local Search

Input: Funzione g, punto iniziale $w^{(1)}$

Inizializzazione: $t = 1$

1. Genera K direzioni casuali $\{d^{(k)}\}_{k=1}^{K}$
2. Valuta $g(w^{(t)} + \eta\, d^{(k)})$ per $k = 1, \ldots, K$
3. Poni $w^{(t+1)} = w^{(t)} + \eta\, d^{(k)}$, dove k è l'indice corrispondente al valore più basso osservato
4. Se $g(w^{(t+1)}) < g(w^{(t)})$, incrementa t e ricomincia

Esempio


minimum point
$g(\mathbf{w})$
candidate direction
chosen direction
$\mathbf{w}^0$
$\mathbf{w}^1$
$\mathbf{w}^2$

Metodi di ordine 0, 1, 2, ...

Randomized Local Search è un esempio di metodo di ordine 0

Un *metodo di ordine 0* utilizza solo i valori della funzione g

Un *metodo di ordine 1* utilizza, in più, i valori delle derivate prime di g

Un *metodo di ordine 2* utilizza, in più, i valori delle derivate seconde di g

Approssimazioni di Taylor: Caso univariato ($N = 1$)

Approssimazione di ordine 1

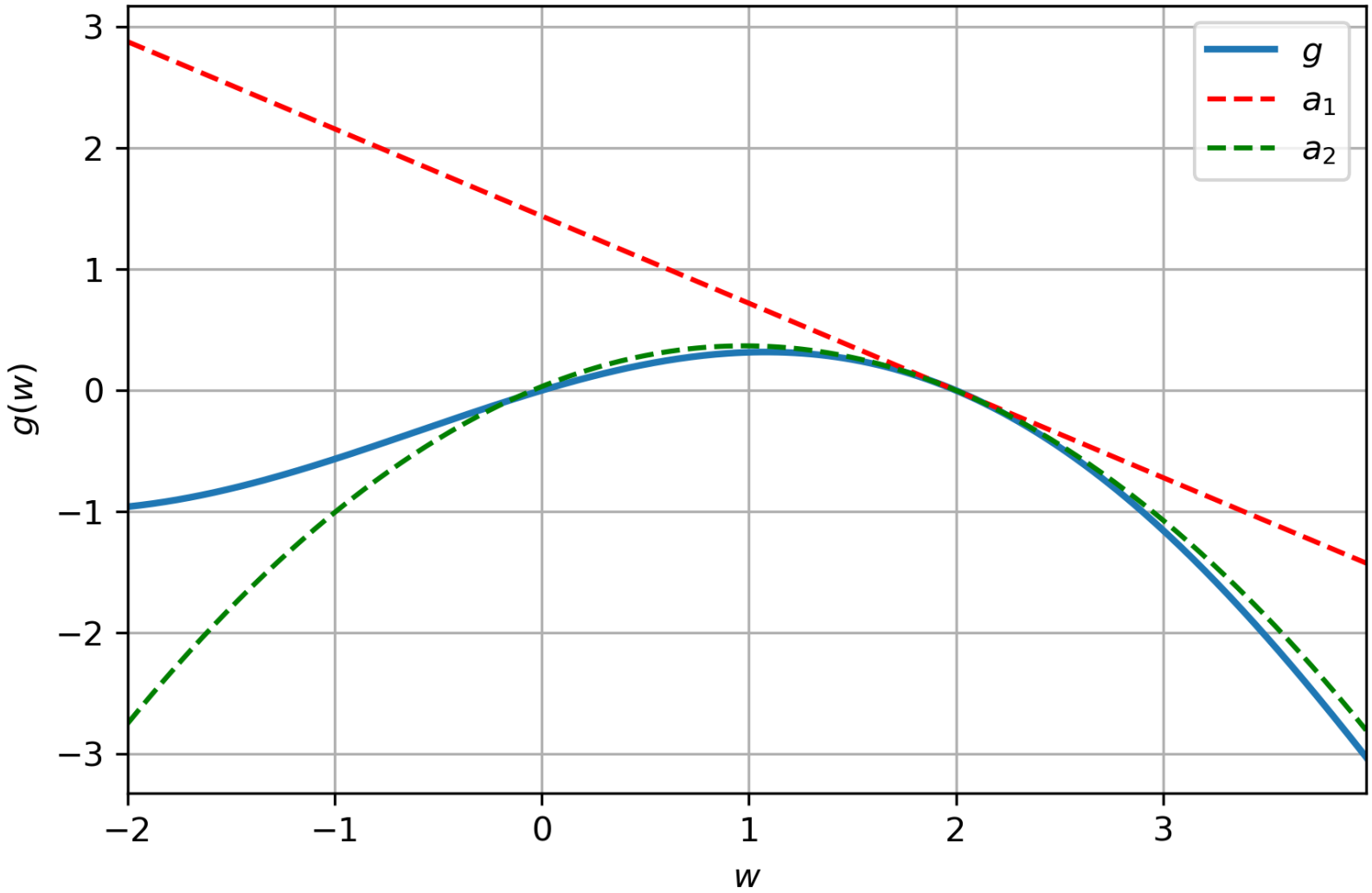
$$a(w) = g(v) + g'(v)(w - v)$$

Approssimazione di ordine 2

$$a(w) = g(v) + g'(v)(w - v) + \frac{1}{2}g''(v)(w - v)^2$$

Esempio

$g(w) = (w + 8)(w + 4)w(w - 2)(w - 8)/1000,\ v = 2$



Gradiente ed Hessiano

Gradiente di g nel punto v

$$\nabla_w g(v) \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial g}{\partial w_1}(v) \\ \frac{\partial g}{\partial w_2}(v) \\ \dots \\ \frac{\partial g}{\partial w_N}(v) \end{bmatrix}$$

Hessiano di g nel punto v

$$\nabla^2_w g(v) \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial^2 g}{\partial w_1^2}(v) & \frac{\partial^2 g}{\partial w_1 \partial w_2}(v) & \dots & \frac{\partial^2 g}{\partial w_1 \partial w_N}(v) \\ \frac{\partial^2 g}{\partial w_2 \partial w_1}(v) & \frac{\partial^2 g}{\partial w_2^2}(v) & \dots & \frac{\partial^2 g}{\partial w_2 \partial w_N}(v) \\ \dots & \dots & \dots & \dots \\ \frac{\partial^2 g}{\partial w_N \partial w_1}(v) & \frac{\partial^2 g}{\partial w_N \partial w_2}(v) & \dots & \frac{\partial^2 g}{\partial w_N^2}(v) \end{bmatrix}$$

Abbreviati in $\nabla g(v)$, $\nabla^2 g(v)$ se le variabili w sono chiare dal contesto

Approssimazioni di Taylor: Caso generale ($N \geq 1$)

Approssimazione di ordine 1

$$a(w) = g(v) + \nabla g(v)^\top (w - v)$$

Approssimazione di ordine 2

$$a(w) = g(v) + \nabla g(v)^\top (w - v) + \frac{1}{2}(w - v)^\top \nabla^2 g(v)(w - v)$$

$\nabla g(v)$ è il *gradiente* di g (vettore delle derivate prime) in v

$\nabla^2 g(v)$ è l'*Hessiano* di g (matrice delle derivate seconde) in v

Condizione di ottimalità al prim'ordine

$$\nabla g(v) = 0$$

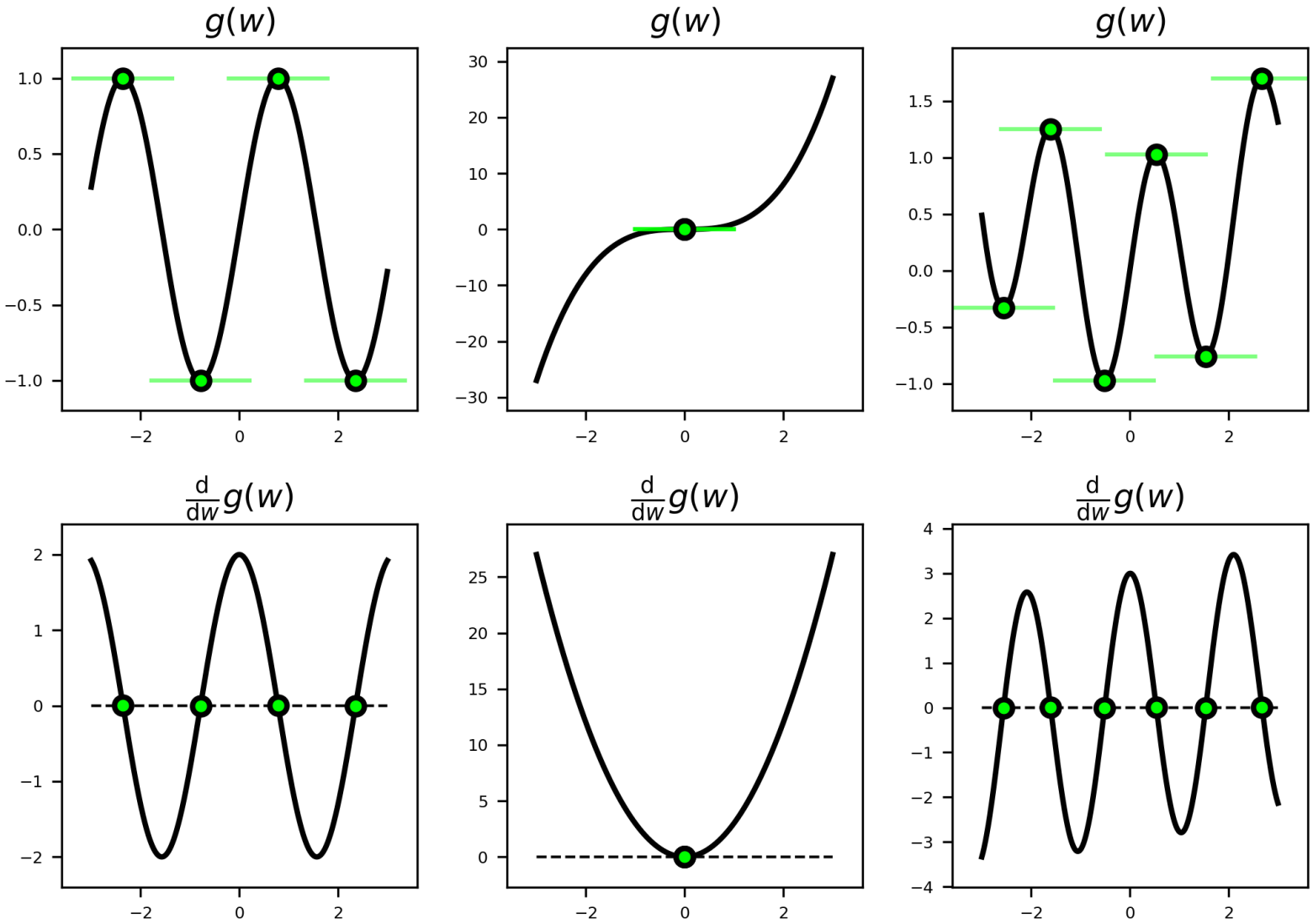
è una condizione necessaria affinché v sia un minimo globale di g

Non è (in generale) sufficiente: identifica solo i *punti critici* di g:

- minimi/massimi locali
- punti di sella

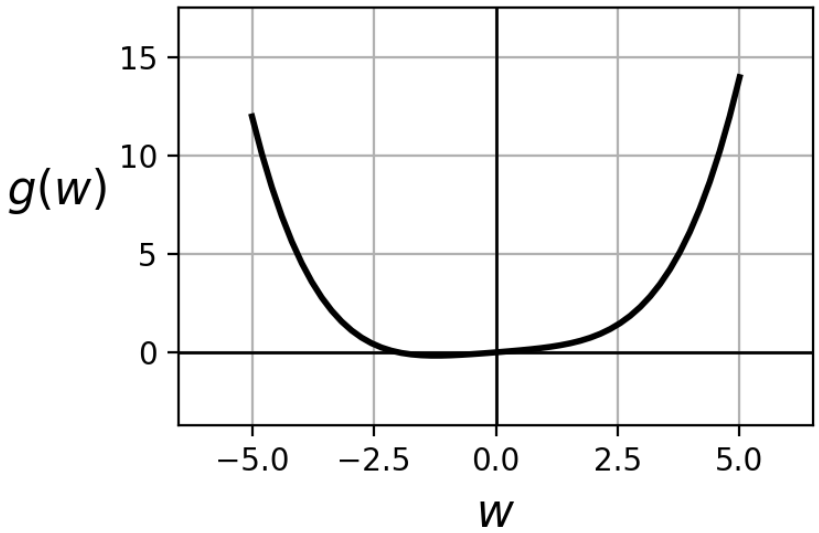
Anche detti *punti stazionari* della funzione

Punti critici: Esempi



Un esempio non banale

$$g(w) = \frac{1}{50}(w^4 + w^2 + 10w)$$



Un altro esempio

$$g(w) = a + b^\top w + w^\top C w$$

con $a \in \mathbb{R}$, $b \in \mathbb{R}^N$, $C \in \mathbb{R}^{N \times N}$ (C simmetrica)

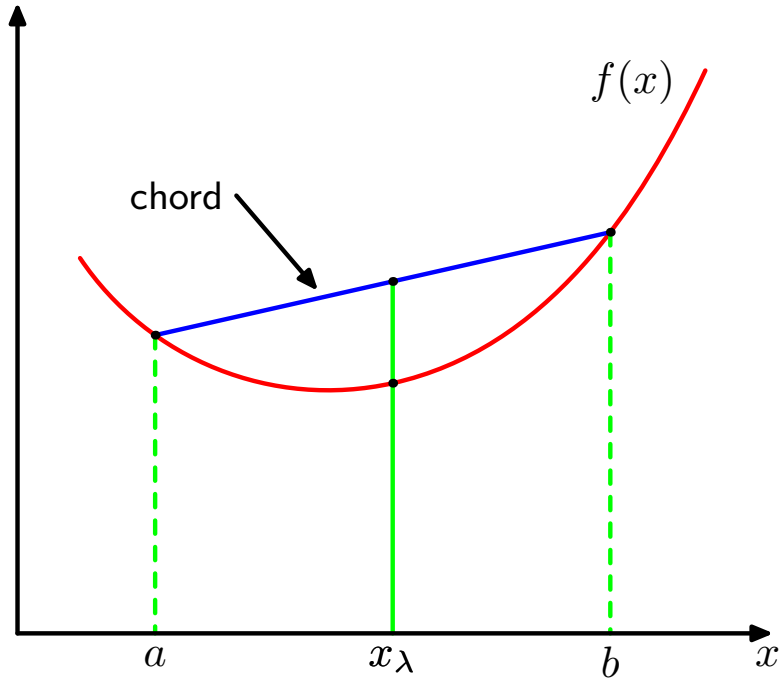
$$\nabla g(w) = 0 \Leftrightarrow Cw = -\frac{1}{2} b$$

Funzioni convesse

Funzione convessa (definizione di ordine 0)

Una funzione $f : \mathbb{R}^N \to \mathbb{R}$ è *convessa* se per ogni $\lambda \in [0, 1]$, $a, b \in \mathbb{R}^N$,

$$f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)f(b)$$



Funzioni convesse: definizione alternativa (ordine 1)

Se f è differenziabile, f è convessa se e solo se

$$f(b) \geq f(a) + \nabla f(a)^\top (b - a) \qquad \text{per ogni } a, b \in \mathbb{R}^N$$

Funzioni convesse: definizione alternativa (ordine 2)

Se f è due volte differenziabile, f è convessa se e solo se

$$\nabla^2 f(x)$$

ha tutti gli autovalori non-negativi, per ogni $x \in \mathbb{R}^N$

Funzioni convesse vs. non convesse: Esempi

Esempi con $N = 1$:

- $g(w) = w^3$
  non è convessa
- $g(w) = e^w$
  è convessa
- $g(w) = \sin(w)$
  non è convessa
- $g(w) = w^2$
  è convessa
- $g(w) = |w|$
  è convessa

Funzioni convesse vs. non convesse: Esempi

Esempi con $N > 1$:

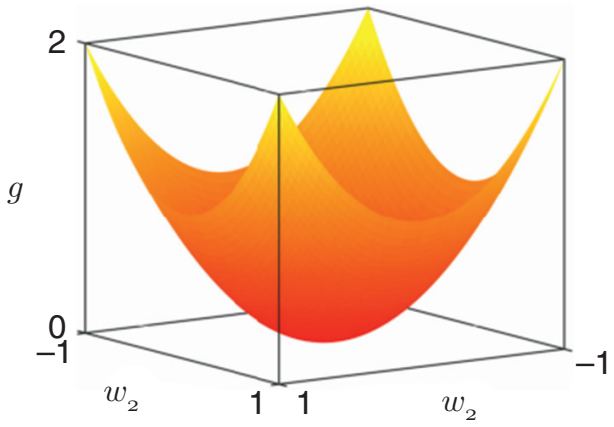
$$g(w) = \frac{1}{2} w^\top Q w + r^\top w + b$$

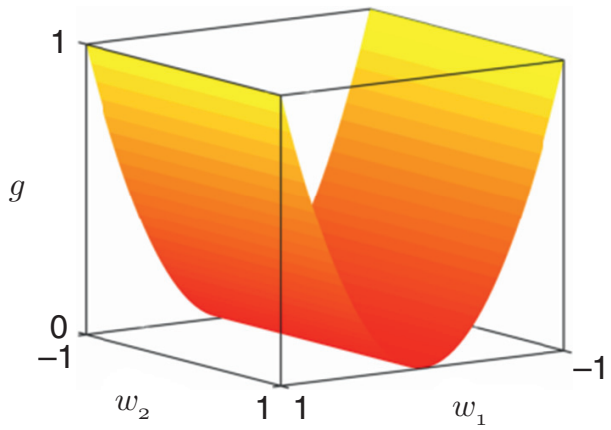
con Q simmetrica

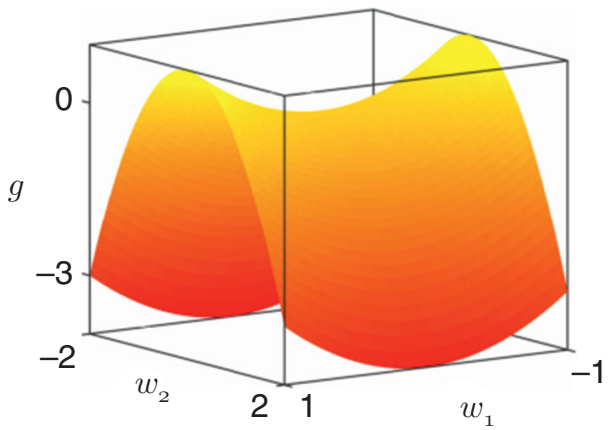
$$\nabla^2 g(w) = Q$$

- con $Q = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ è convessa
- con $Q = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$ è convessa
- con $Q = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$ non è convessa

Funzioni convesse vs. non convesse: Esempi







Alcuni criteri sufficienti di convessità

- Ogni funzione lineare è convessa
- Se f è convessa e $c \geq 0$, $c \cdot f(x)$ è convessa
- Se f e g sono convesse, $f(x) + g(x)$ è convessa
- Se f e g sono convesse, $\max(f(x), g(x))$ è convessa
- Se a è lineare e f è convessa, $f(a(x))$ è convessa
- Se C è simmetrica con autovalori ≥ 0, $x^\top Cx$ è convessa
- Se $C = v \cdot v^\top$, $x^\top Cx$ è convessa
- Se $\nabla^2 f$ è simmetrica con autovalori ≥ 0, $f(x)$ è convessa

Norme

Una *norma* è una funzione $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$ con le seguenti proprietà:

1. $\|x\| = 0$ se e solo se $x = 0_{N\times 1}$
2. $\|x\| \geq 0$ per ogni $x \in \mathbb{R}^N$
3. $\|cx\| = c\,\|x\|$ per ogni $x \in \mathbb{R}^N$ e $c > 0$
4. $\|x + y\| \leq \|x\| + \|y\|$

Esempi:

- norma euclidea: $\ell_2(x) = (\sum_j |x_j|^2)^{1/2}$
- norma ℓ_1: $\ell_1(x) = \sum_j |x_j|$
- norma ℓ_p: $\ell_p(x) = (\sum_j |x_j|^p)^{1/p}$ per qualche $p \geq 1$
- norma uniforme: $\ell_\infty(x) = \max_j |x_j|$

Ogni norma è una funzione convessa (perché?)

Condizione di ottimalità per funzioni convesse

Se g è convessa e differenziabile, la condizione

$$\nabla g(w) = 0$$

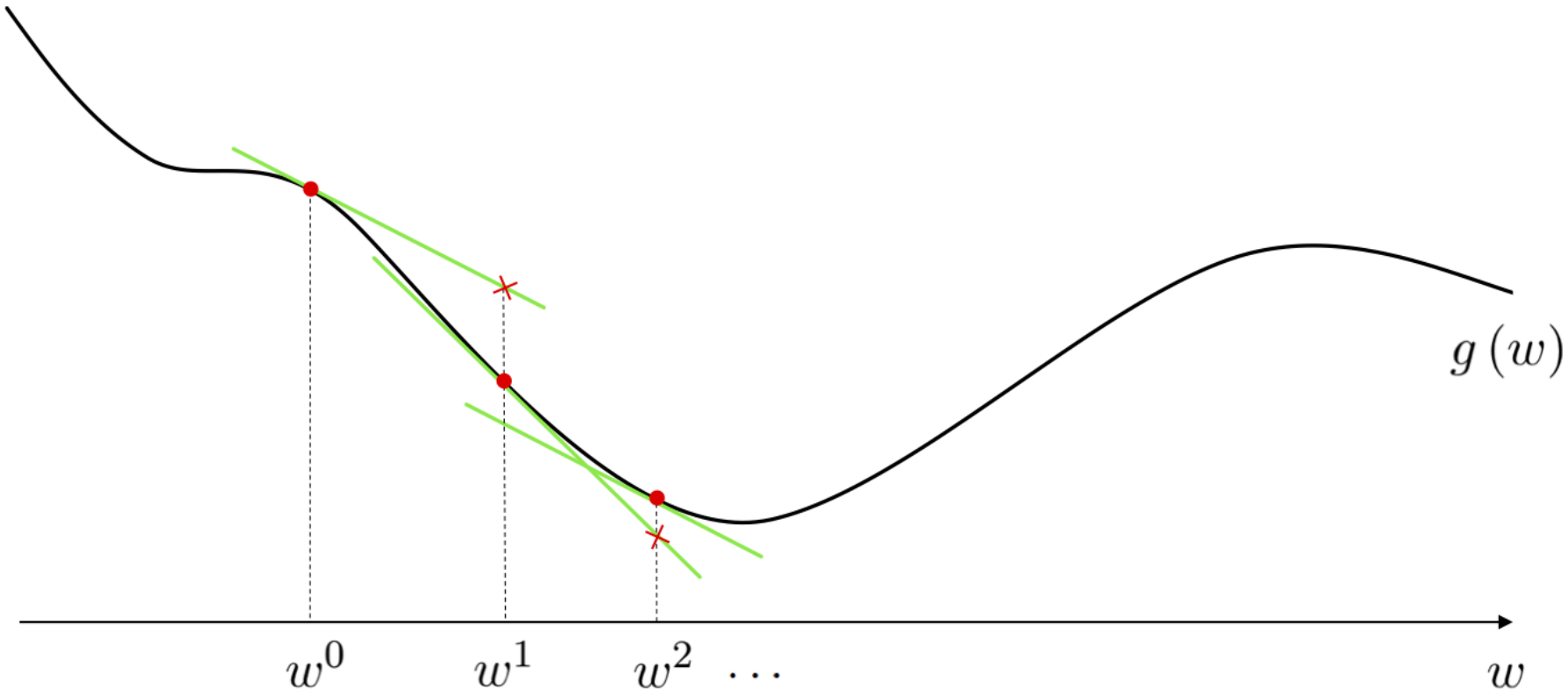
è necessaria e sufficiente affinché w sia un minimo globale.

Infatti, per ogni w',

$$g(w') \geq g(w) + \nabla g(w)^\top (w' - w) = g(w)$$

$\Rightarrow$ Nelle funzioni convesse, i minimi locali sono anche globali

Discesa del gradiente [Gradient Descent]



Algoritmo di discesa del gradiente (GD)

Algoritmo Gradient Descent (versione generica)

Input: Funzione g, punto iniziale $w^{(1)}$

1. Per $t = 1, \ldots, T$:

$$w^{(t+1)} = w^{(t)} - \eta \cdot \nabla g(w^{(t)})$$

2. Restituisci il $w^{(t)}$ col minimo valore di $g(w^{(t)})$, $t = 1, \ldots, T$

L'algoritmo ha due parametri: η (passo) e T (numero di passi) (chiamati *iperparametri* per non confonderli con le variabili w da ottimizzare)

Convergenza di GD

Teorema (Convergenza di GD – Versione 1)

Sia $g : \mathbb{R}^N \to \mathbb{R}$ convessa. Se esistono costanti $D, G > 0$ tali che:

- $\left\| w^{(1)} - w^* \right\| \leq D$,
- $\left\| \nabla g(w^{(t)}) \right\| \leq G$ per $t = 1, \ldots, T$,

allora GD con $\eta = D/(G\sqrt{T})$ soddisfa

$$g(w^{\mathrm{GD}}) - g(w^*) \leq \frac{DG}{\sqrt{T}}$$

$\Rightarrow$ GD converge al valore minimo di g per T crescente

L'errore decresce asintoticamente almeno come $1/\sqrt{T}$

Convergenza di GD

Teorema (Convergenza di GD – Versione 2)

Sia $g : \mathbb{R}^N \to \mathbb{R}$ convessa e β-smooth. Allora GD con $\eta = 1/\beta$ soddisfa

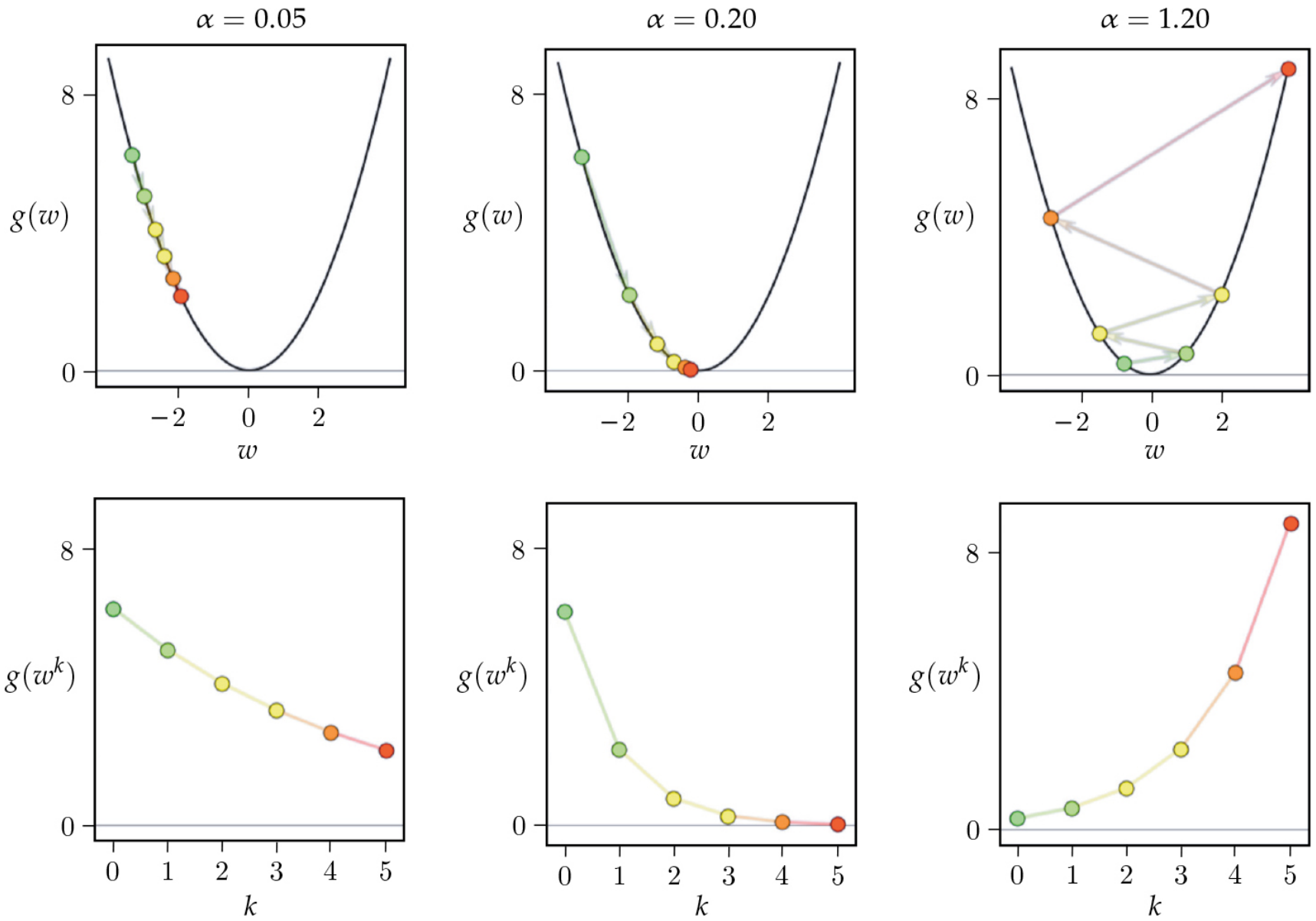
$$g(w^{\mathrm{GD}}) - g(w^*) \leq \frac{2\beta \left\| w^{(1)} - w^* \right\|^2}{T - 1}$$

Nota. Una funzione convessa $g : \mathbb{R}^N \to \mathbb{R}$ è detta *β-smooth* se $\forall x, y$

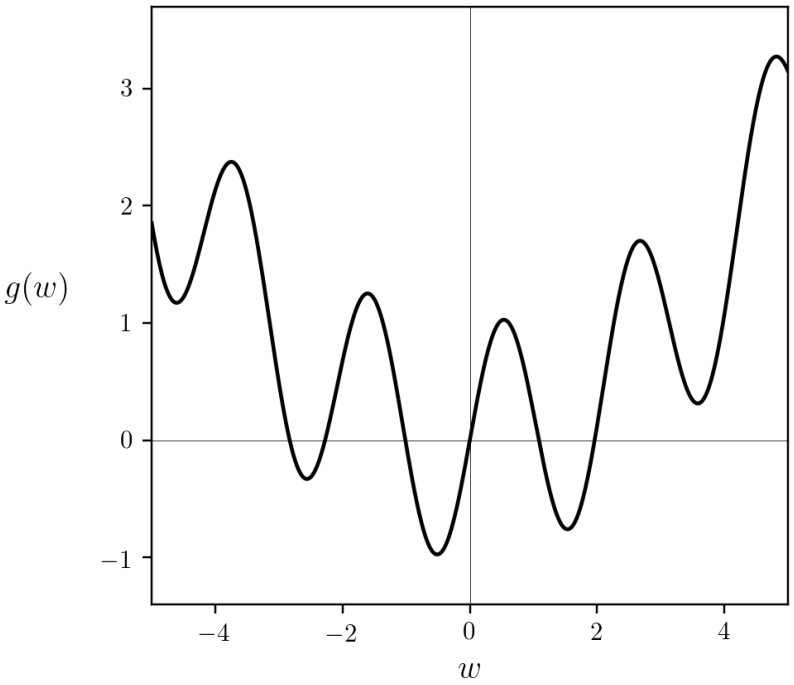
$$g(y) - g(x) - \nabla g(x)^\top (y - x) \leq \frac{\beta}{2} \|x - y\|^2$$

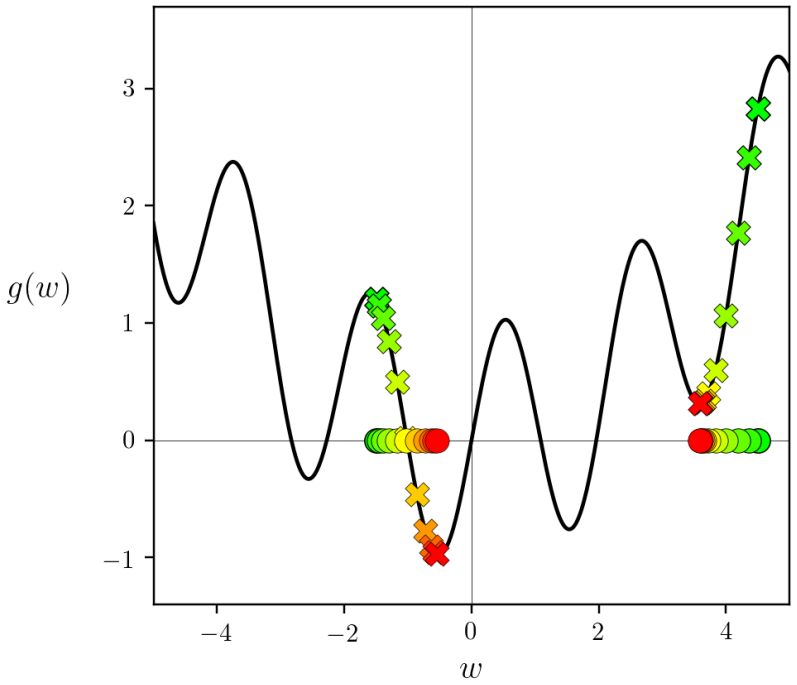
In questo scenario l'errore decresce asintoticamente almeno come $1/T$

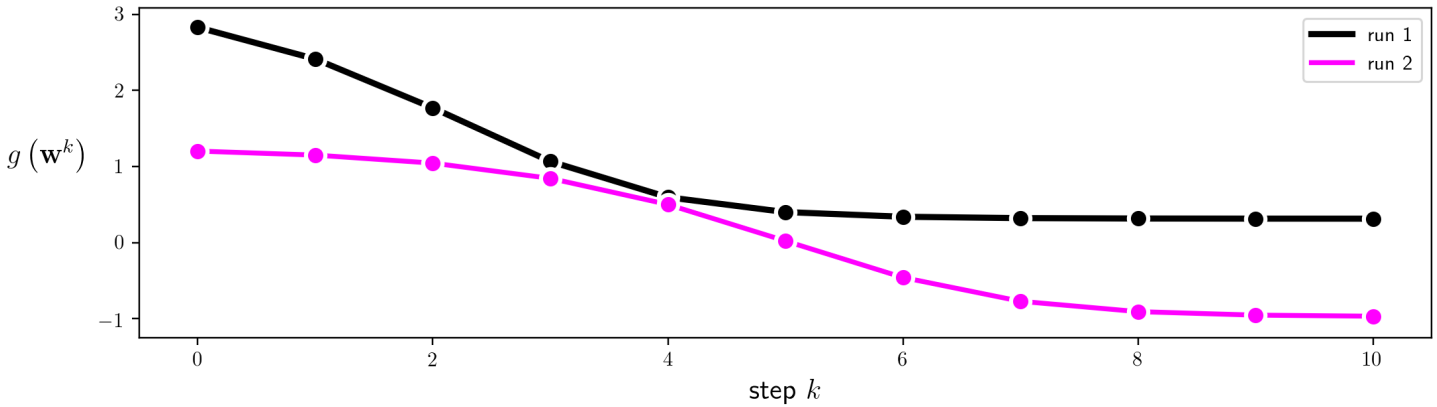
Esempio con diversi valori del passo η



Esempio non convesso



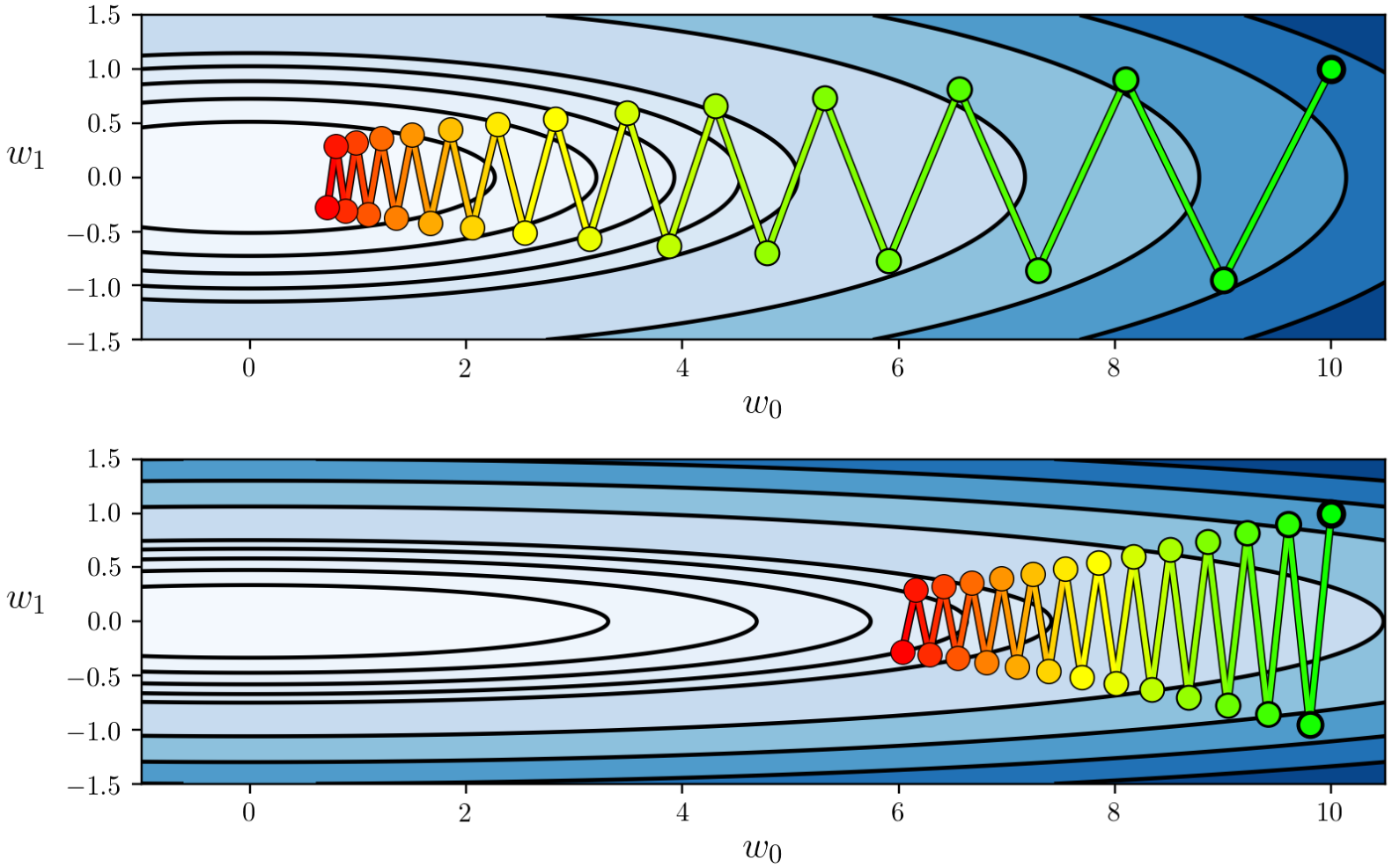




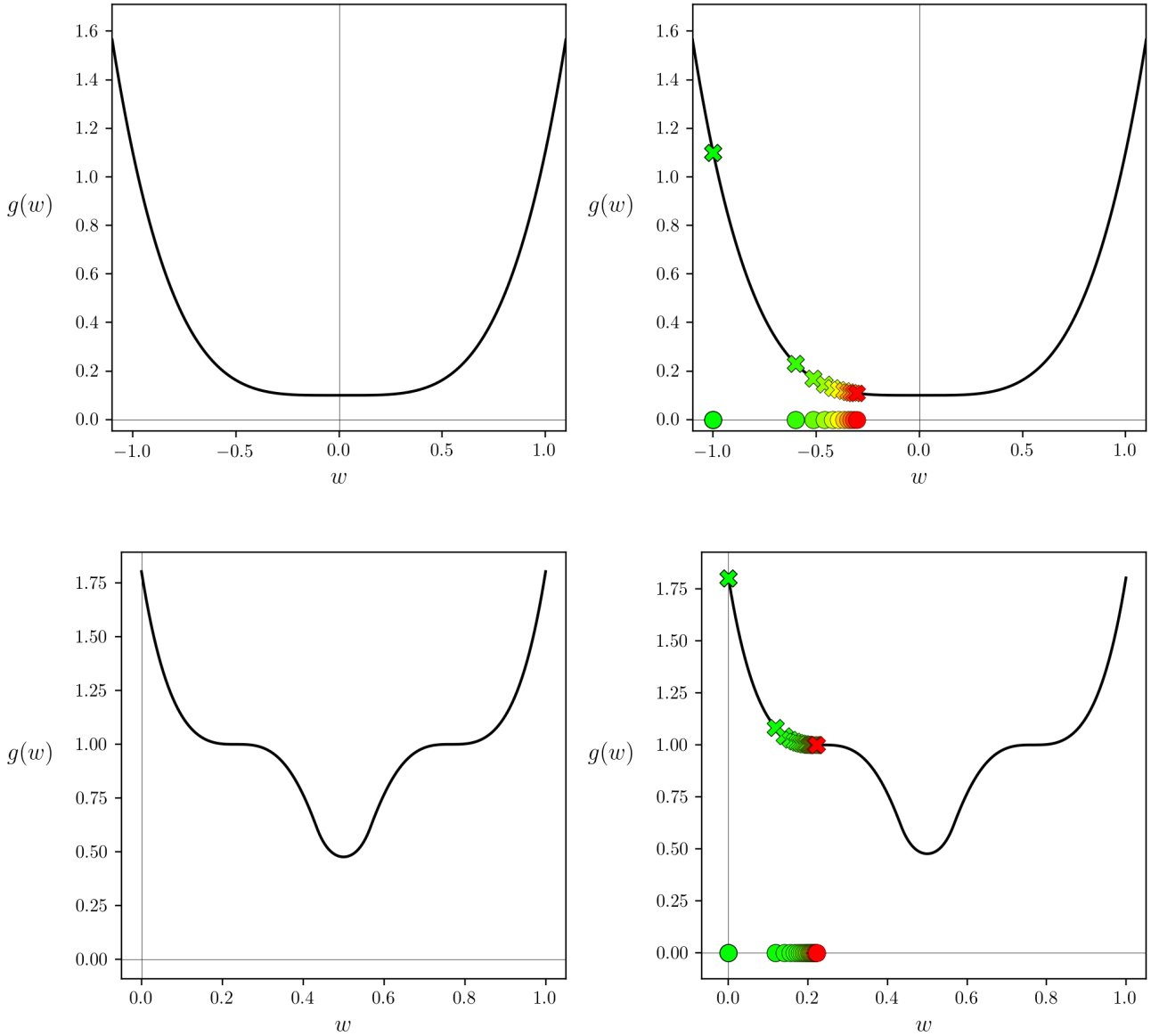
Due problematiche di GD

- La direzione del gradiente negativo può oscillare, portando l'algoritmo a muoversi a "zig-zag" e convergere lentamente
- La magnitudine del gradiente negativo si contrae vicino ai punti critici, rallentando la discesa

Movimento a zig-zag: Esempi

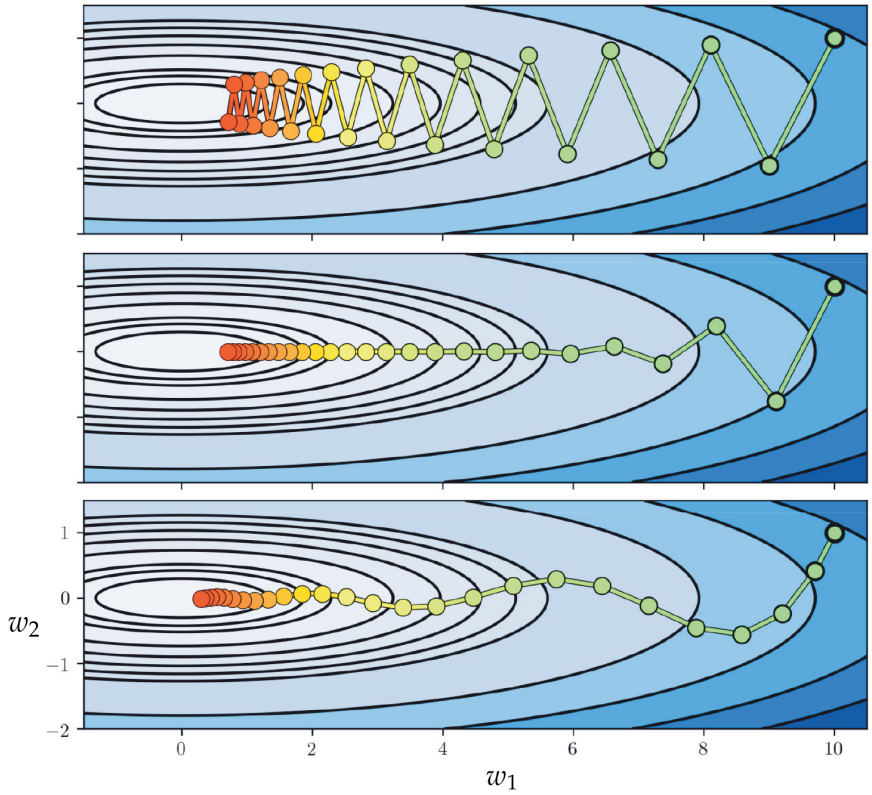


Rallentamento vicino ai punti critici: Esempi



Varianti sofisticate di GD

- Momentum
- Gradiente normalizzato
- Adagrad/RMSprop
- Adam



Discussi nell'Appendice A del libro di testo

Metodi del secondo ordine

GD è un esempio di *metodo del primo ordine* in quanto utilizza solo:

- i valori della funzione, $g(x)$ (scalari)
- i valori del gradiente, $\nabla g(x)$ (vettori $N \times 1$)

I *metodi del secondo ordine* utilizzano anche

- i valori dell'Hessiano, $\nabla^2 g(x)$ (matrici $N \times N$)

Il *metodo di Newton* ne è l'esempio più noto

L'uso di questi metodi nel ML è limitato dal fatto che richiedono la manipolazione esplicita di matrici $N \times N$ (ad ogni passo dell'algoritmo)

Metodo di Newton

Metodo di Newton-Raphson

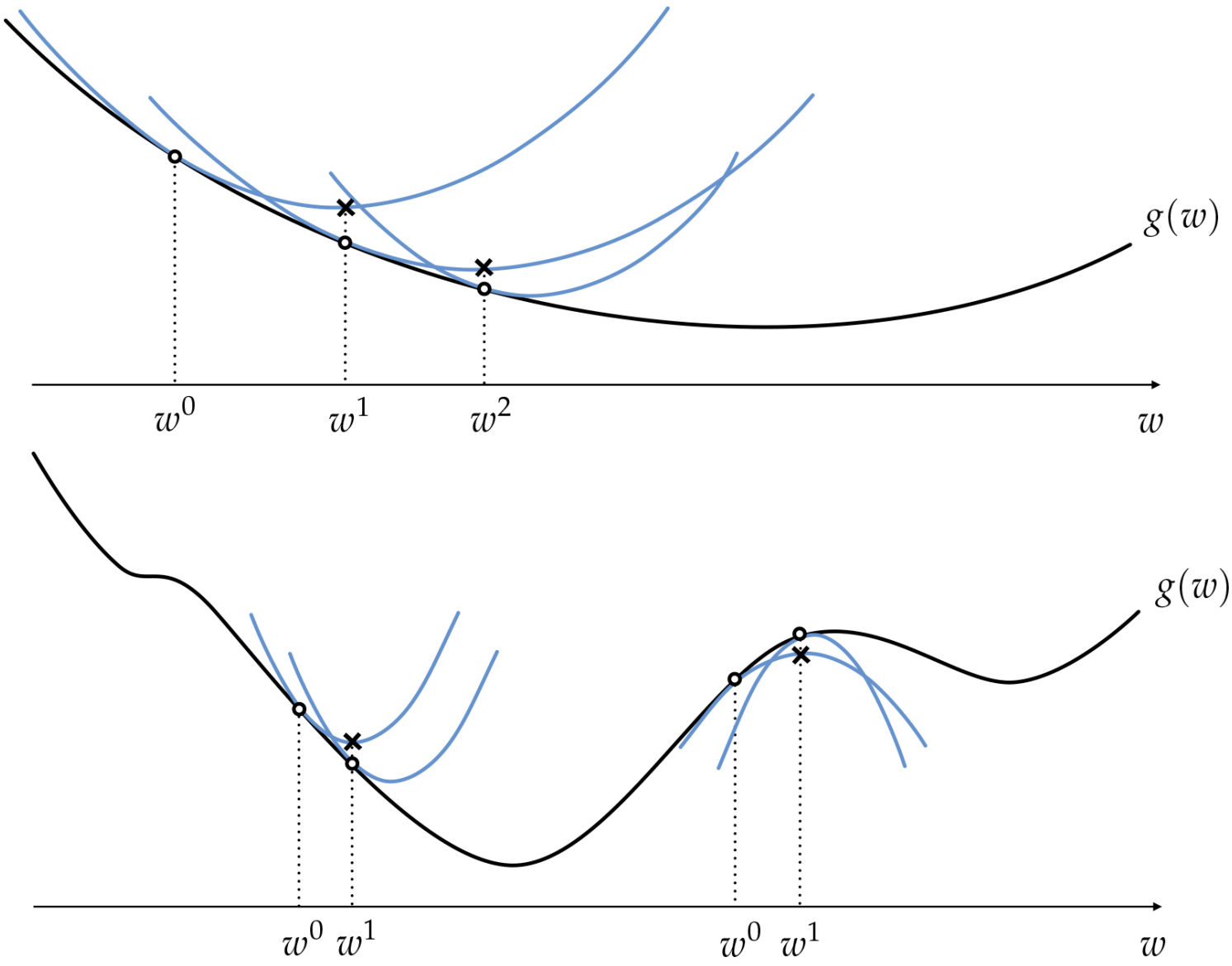
Input: Funzione g, punto iniziale $w^{(1)}$

1. Per $t = 1, \ldots, T$:

$$w^{(t+1)} = w^{(t)} - \left[\nabla^2 g(w^{(t)})\right]^{-1} \nabla g(w^{(t)})$$

2. Restituisci $w^{(T)}$

Esempio



Vantaggi e svantaggi dei metodi del secondo ordine

- Richiedono tipicamente meno iterazioni per convergere
- La singola iterazione è più computazionalmente costosa (richiede di invertire un Hessiano $N \times N$)
- Applicabili solo a funzioni differenziabili e convesse (altrimenti l'Hessiano non è definito, o non è sempre invertibile)