Introduzione al Machine Learning: Metodi di ottimizzazione

Vincenzo Bonifaci

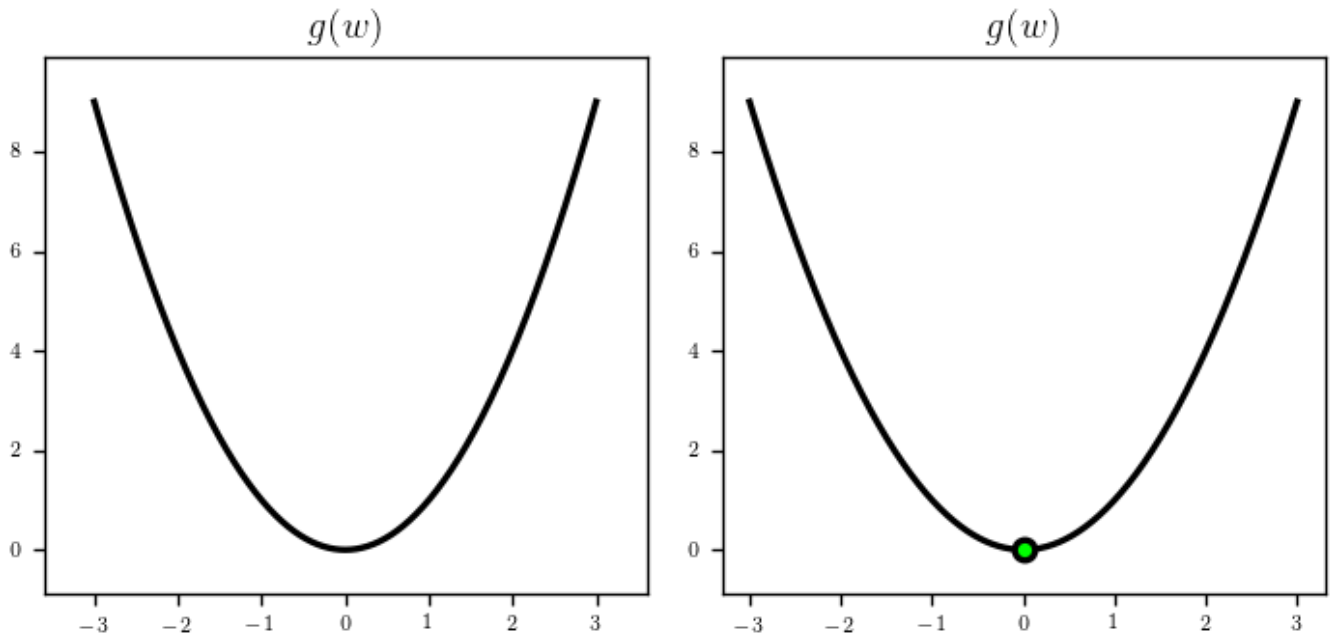
Ottimizzazione matematica

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 $w \in \mathbb{R}^N$

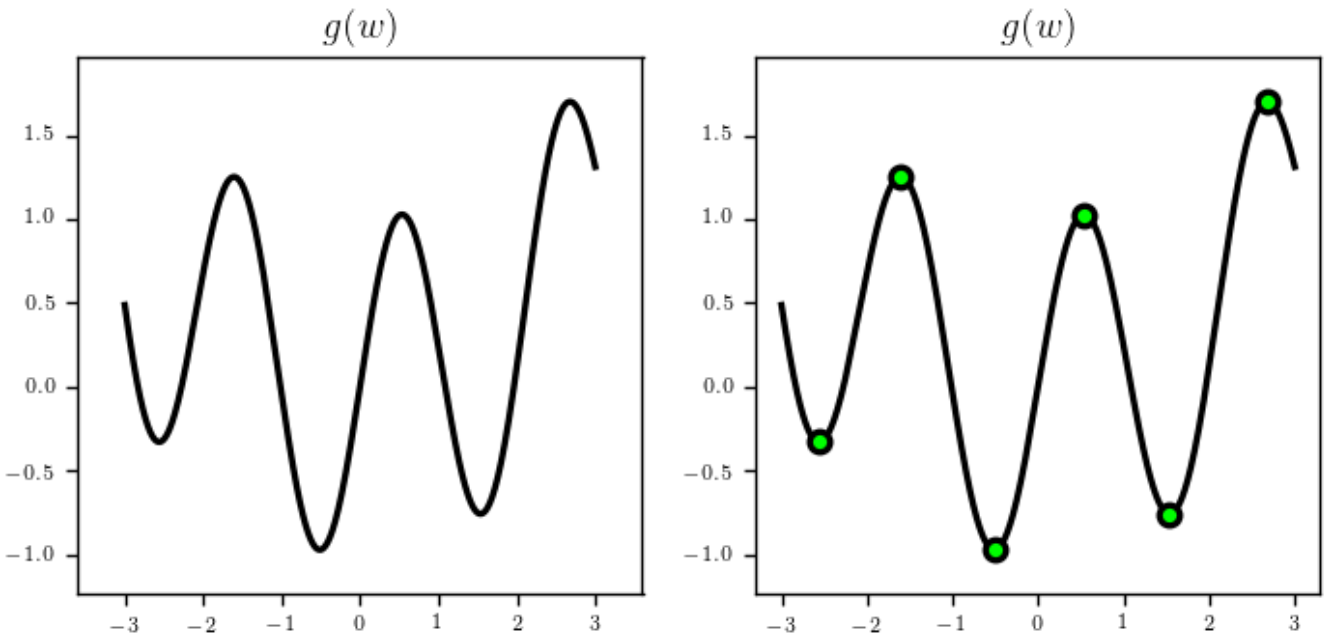
w^* è un *minimo globale* della funzione g

Minimi globali



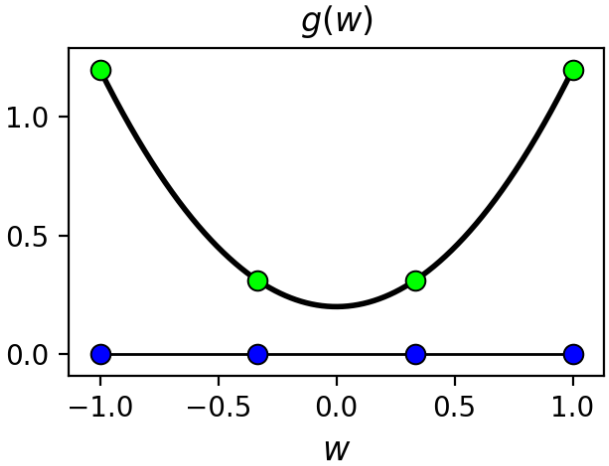
$w^* \in \mathbb{R}^N$ tale che $g(w^*) \leq g(w)$ per ogni w in $\mathbb{R}^N$

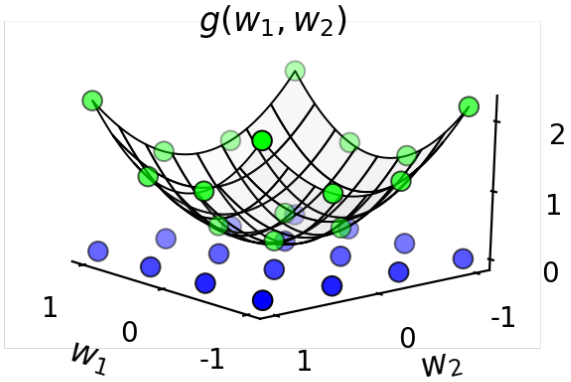
Minimi locali

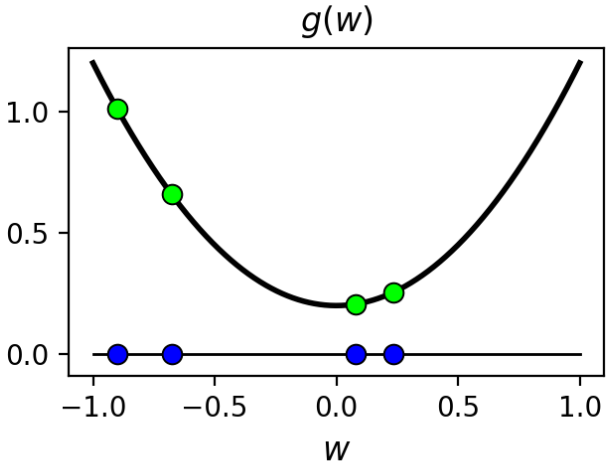


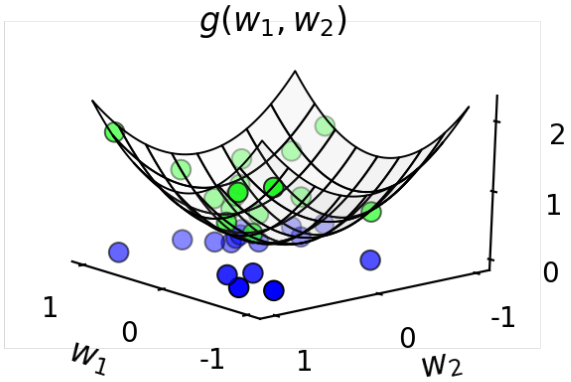
$w^* \in \mathbb{R}^N$ tale che $g(w^*) \leq g(w)$ per ogni w in un intorno di w^*

Due semplici metodi di approssimazione

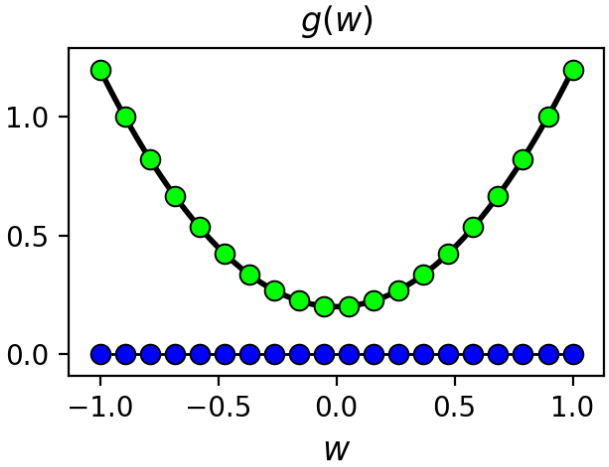


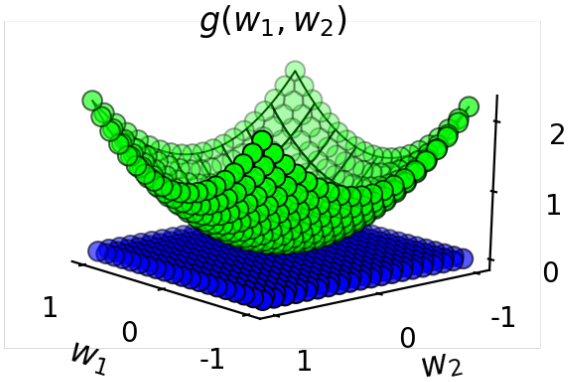


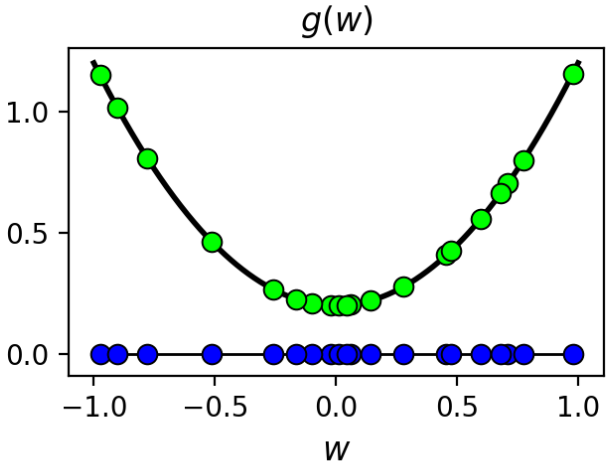


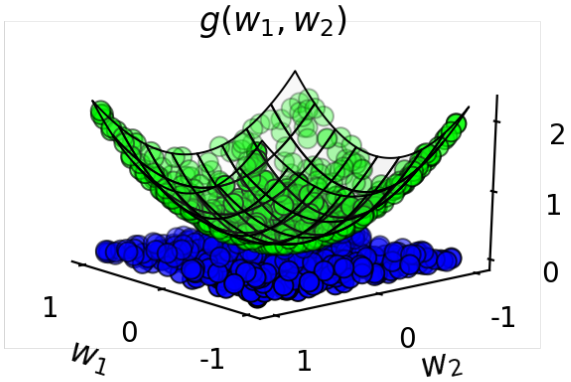


Due semplici metodi di approssimazione

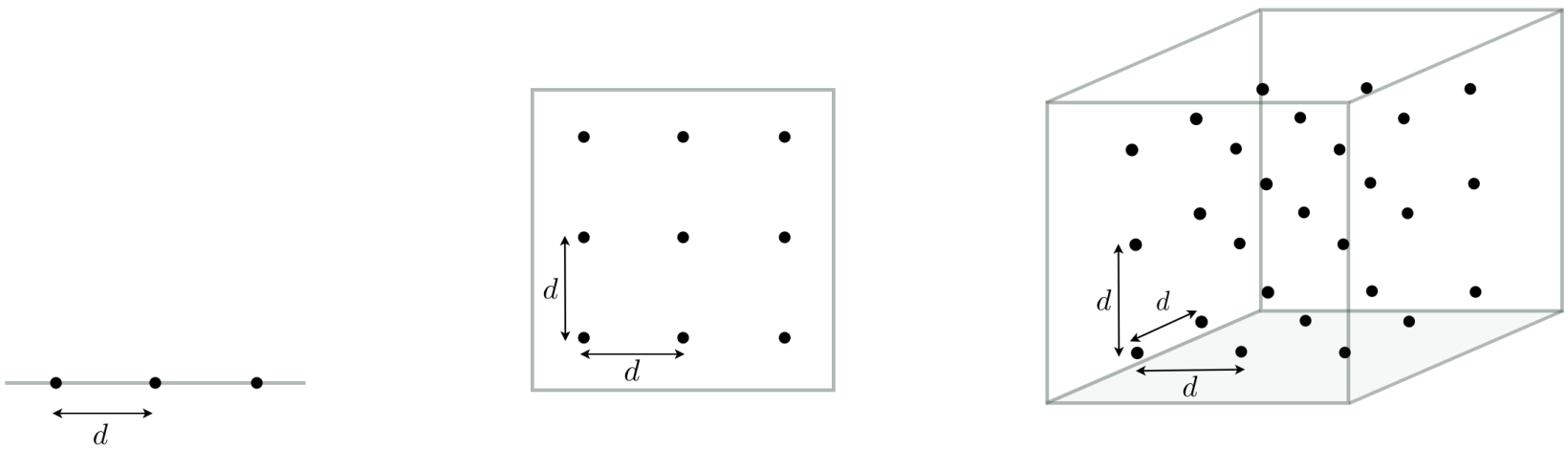






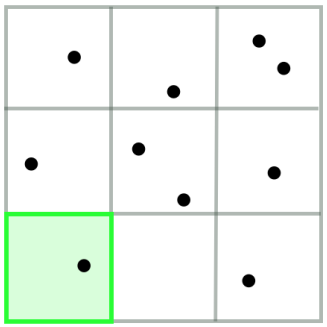


La maledizione della multidimensionalità

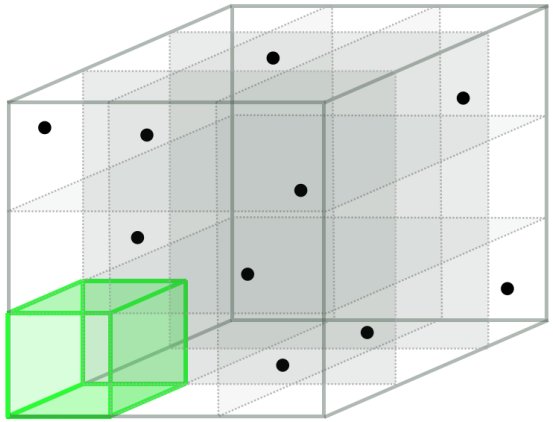


La maledizione della multidimensionalità

$3/10$

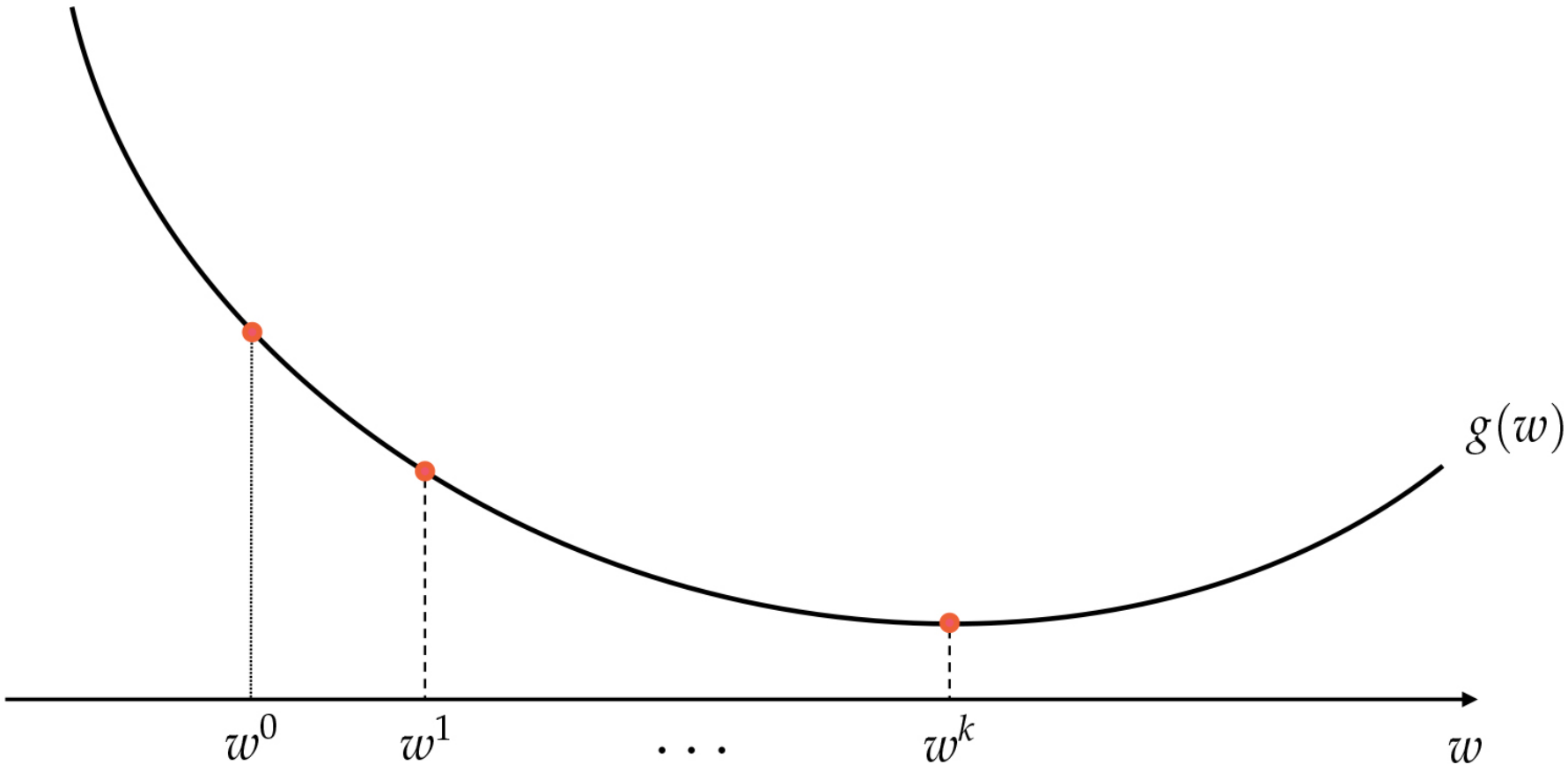


$1/10$

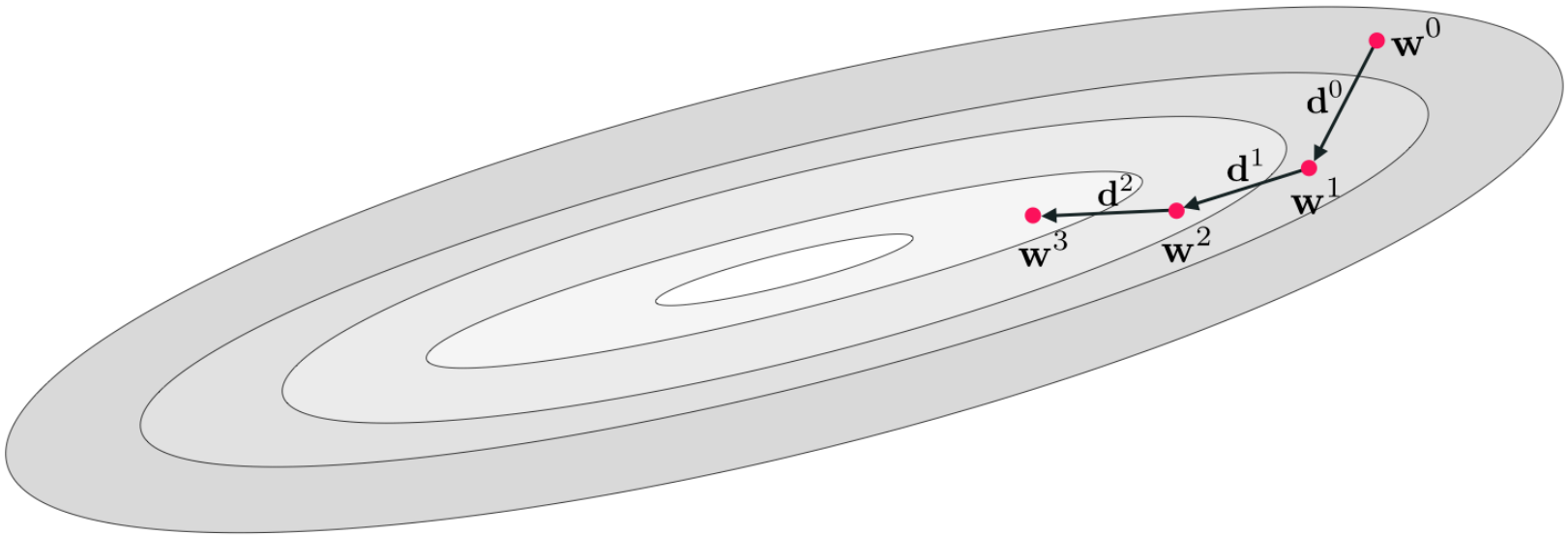


$0/10$

Metodi di ricerca locale



Metodi di ricerca locale

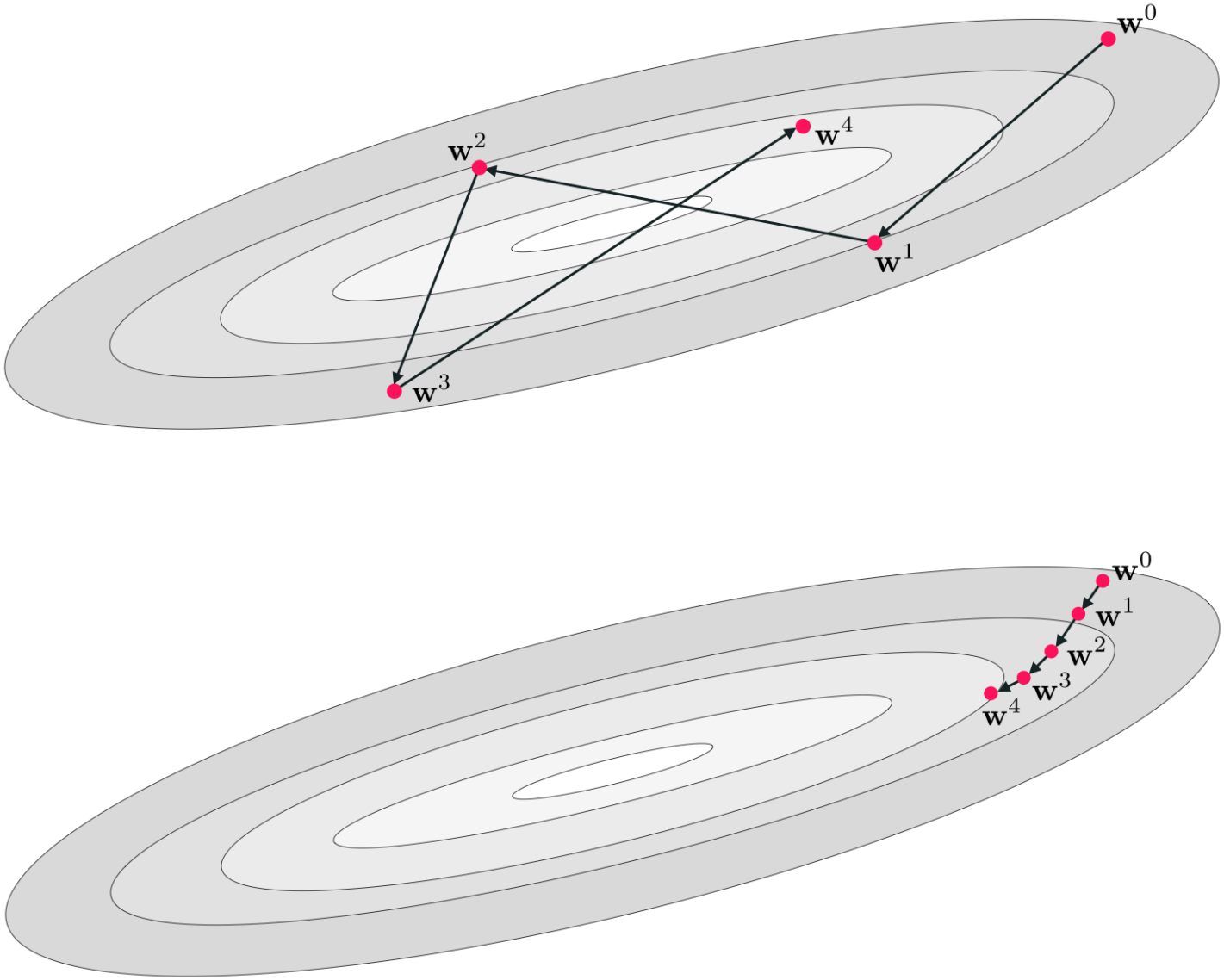


$$w^{(t+1)} = w^{(t)} + d^{(t)}$$

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 *passo* (o *tasso di apprendimento*)

Poiché

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

la lunghezza dello spostamento è direttamente proporzionale a η

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

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

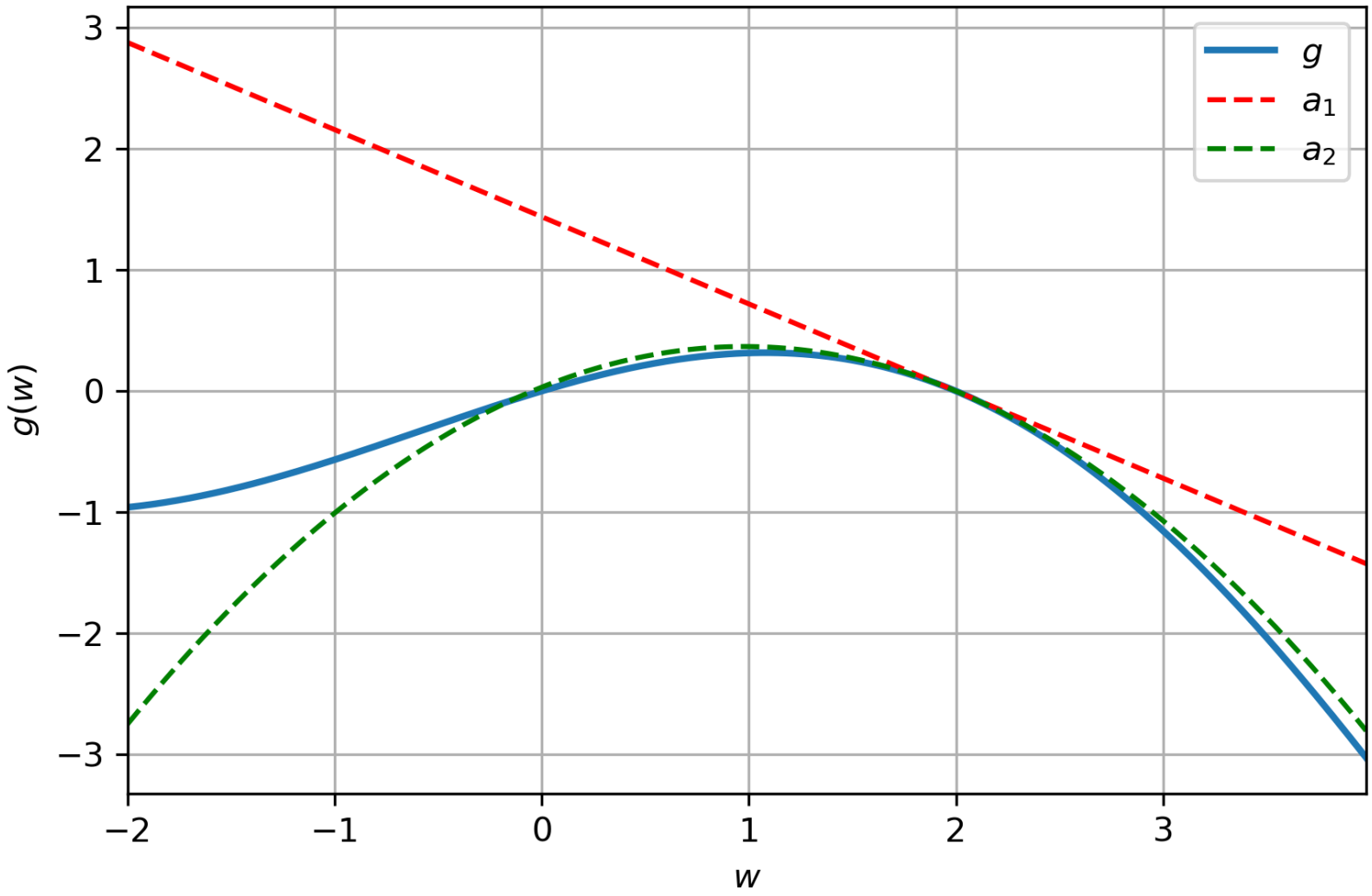
$$a_1(w) = g(v) + g'(v)(w - v)$$

Approssimazione di ordine 2

$$a_2(w) = g(v) + g'(v)(w - v) + \frac{1}{2}g''(v)(w - v)^2$$

Esempio

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



Gradiente ed Hessiana

Gradiente di g nel punto v

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

Hessiana 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

$$a_1(w) = g(v) + \nabla g(v)^\top (w - v)$$

Approssimazione di ordine 2

$$a_2(w) = g(v) + \nabla g(v)^\top (w - v) + \frac{1}{2}(w - v)^\top \nabla^2 g(v)(w - v)$$

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

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

Condizione di ottimalità al prim'ordine

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

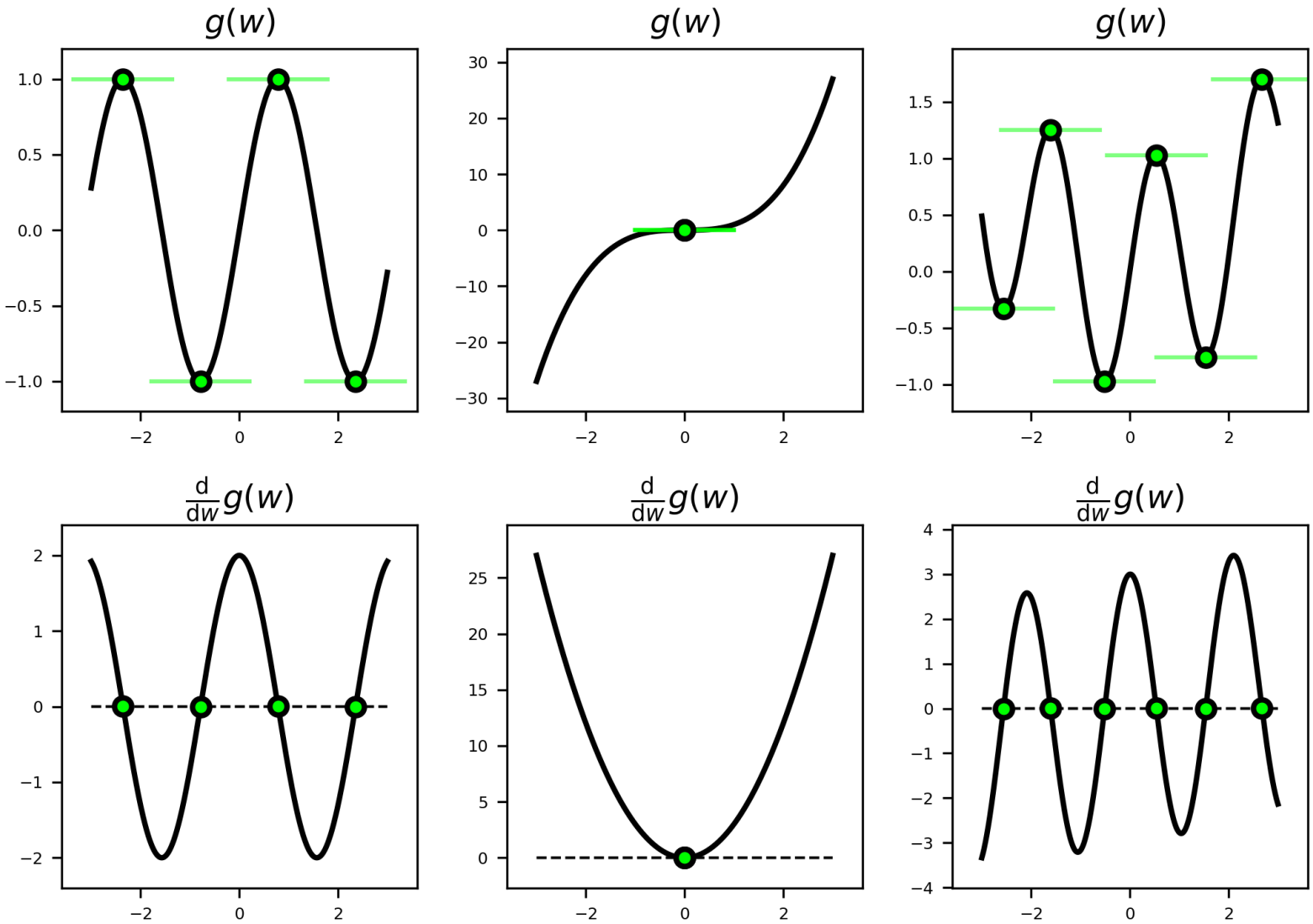
è una condizione necessaria affinché v sia un minimo globale di g

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

- minimi/massimi locali
- punti di sella

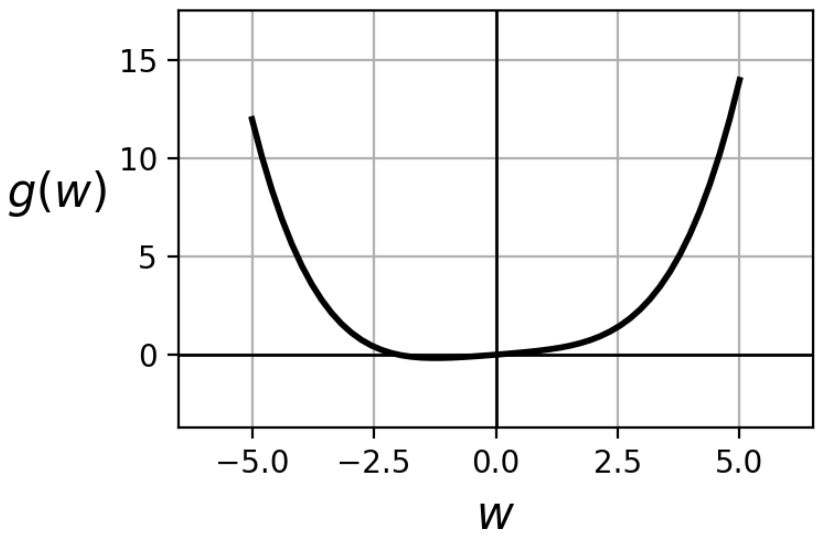
(anche detti *punti stazionari* della funzione)

Punti critici: Esempi



Un esempio

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

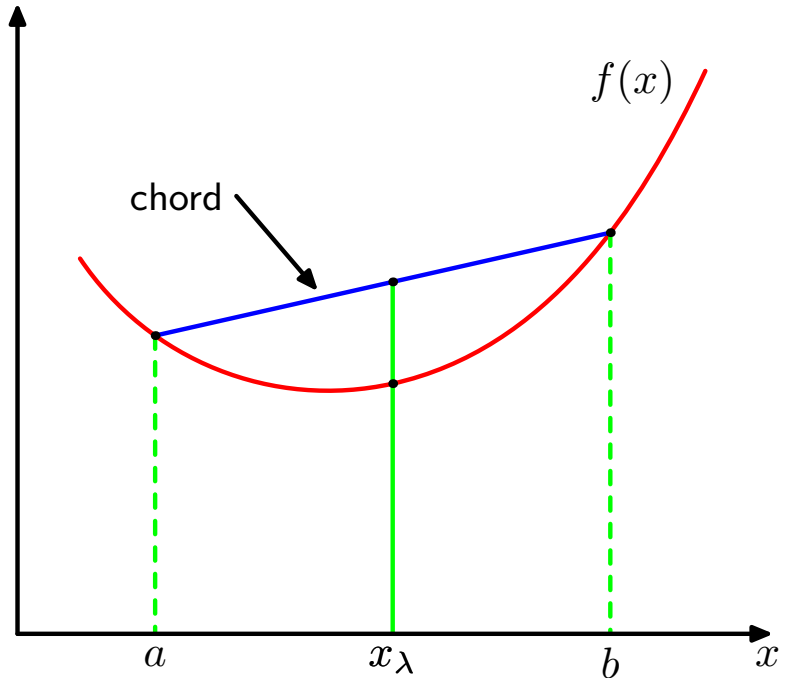


Funzioni convesse

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 l'Hessiana

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

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

Una matrice con tutti gli autovalori ≥ 0 è detta *semidefinita positiva*

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 $N > 1$:

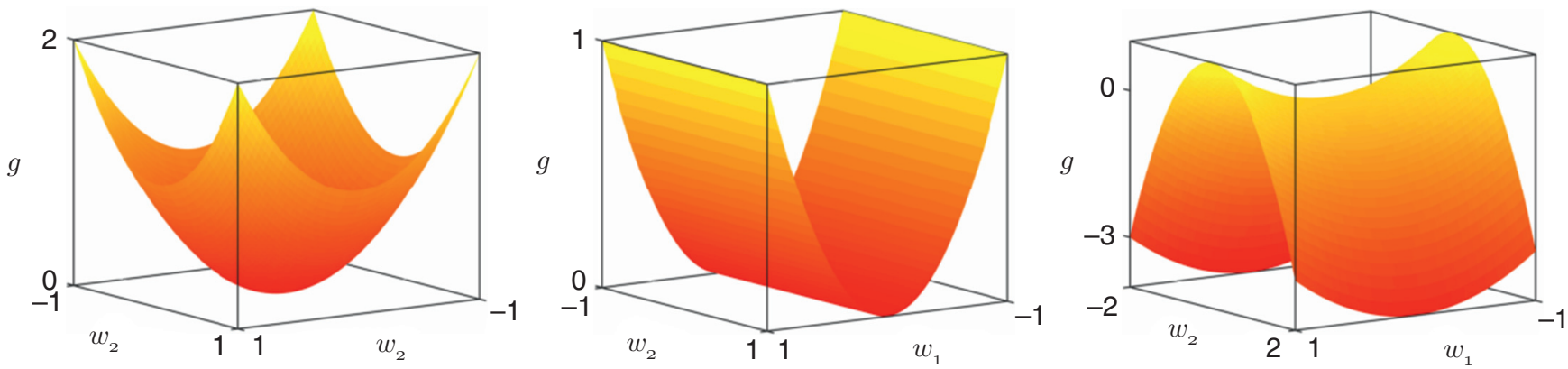
$$g(w) = \frac{1}{2} w^\top Q w + r^\top w + b$$

con Q simmetrica

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

- con $Q = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$ è convessa
- con $Q = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$ è convessa
- con $Q = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}$ non è convessa

Funzioni convesse vs. non convesse: Esempi



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 $\Leftarrow$

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

Convessità e funzioni di costo

- La funzione *quadrato dell'errore*

$$\ell(y) = (y - y^*)^2$$

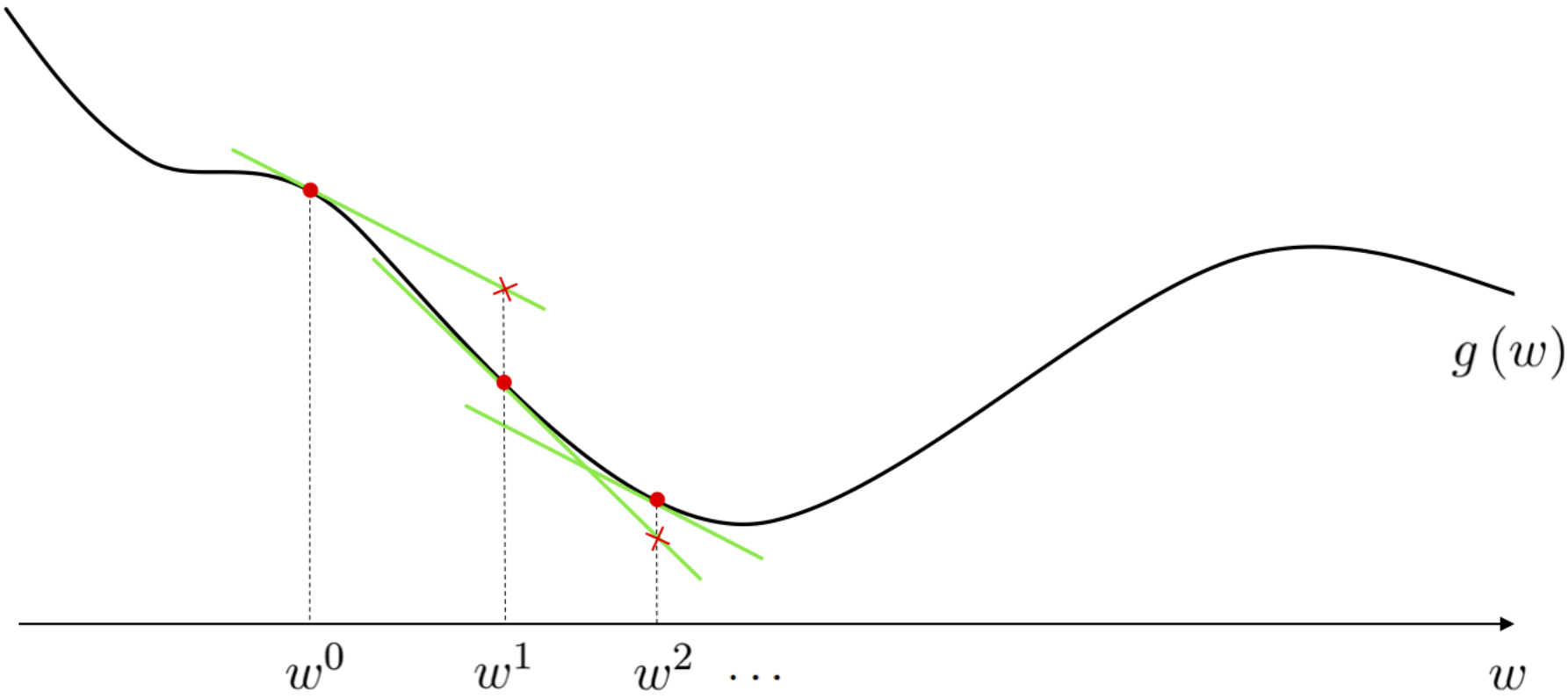
è convessa

- La funzione *costo 0-1*

$$\ell(y) = \begin{cases} 0 & \text{se } y = y^* \\ 1 & \text{se } y \neq y^* \end{cases}$$

non è convessa

Discesa del gradiente [Gradient Descent]



Algoritmo di discesa del gradiente (GD)

Algoritmo Gradient Descent (versione generica)

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

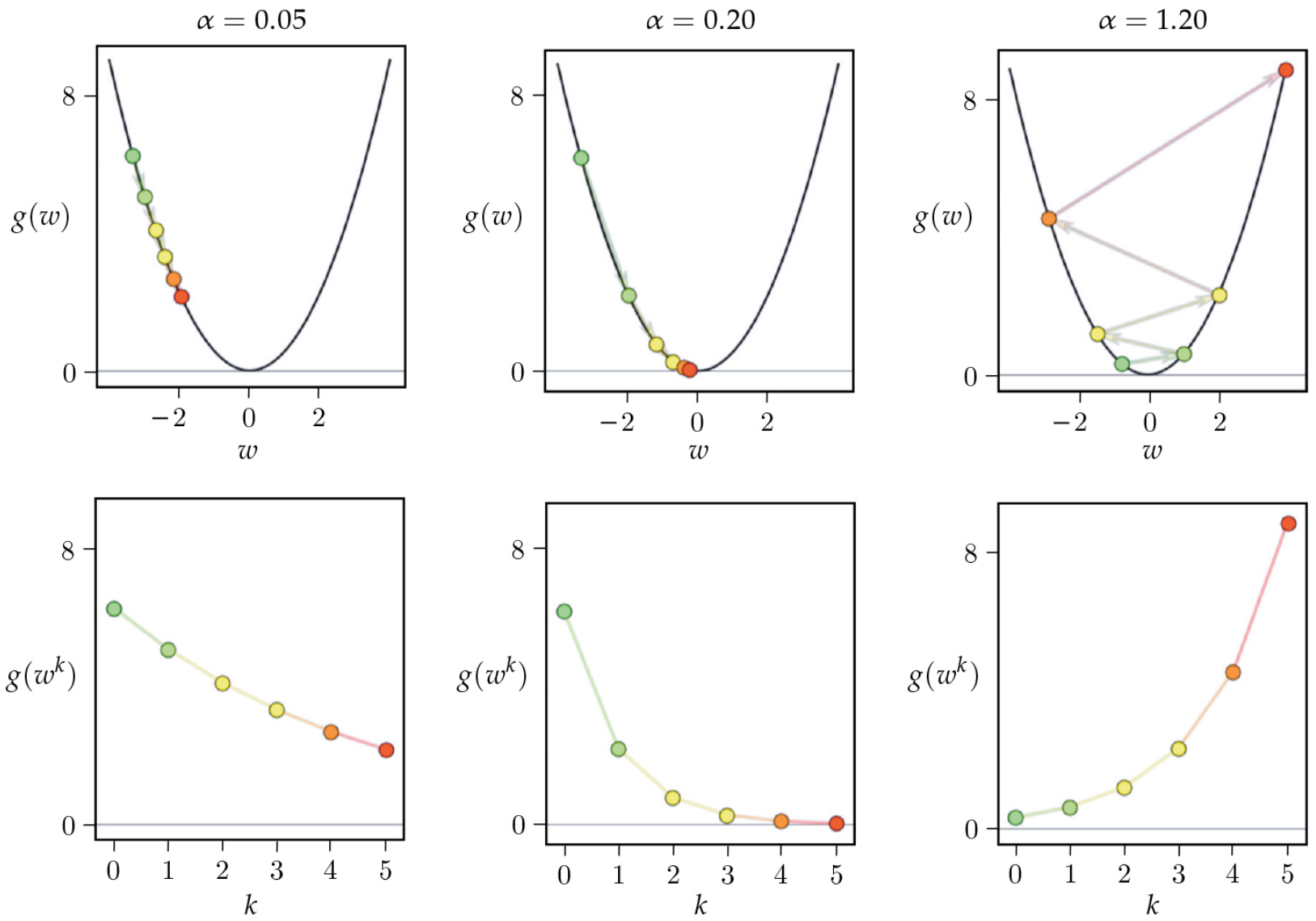
1. Per $t = 1, \ldots, T$:

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

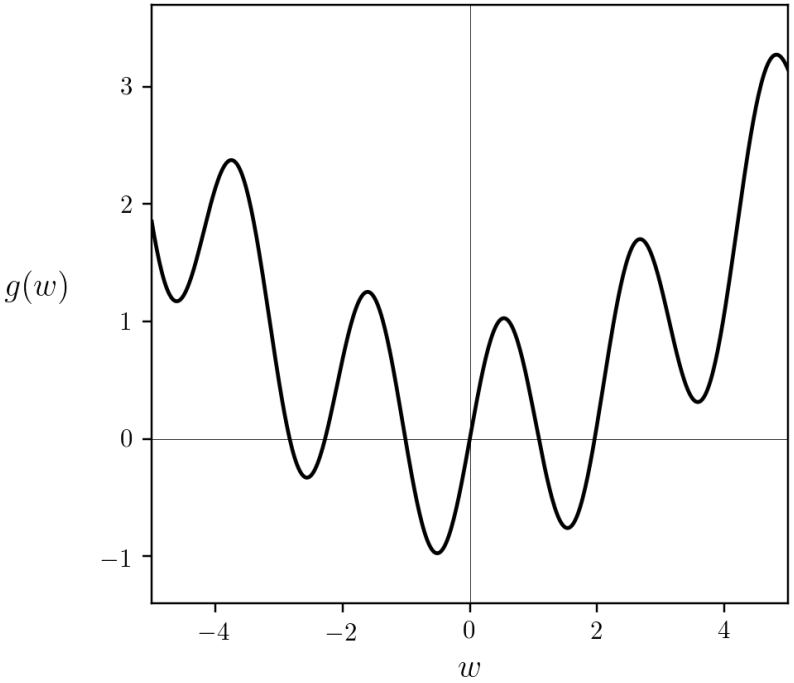
2. Restituisci il $w^{(t)}$ col minimo valore di $g(w^{(t)})$, $t = 1, \ldots, T$

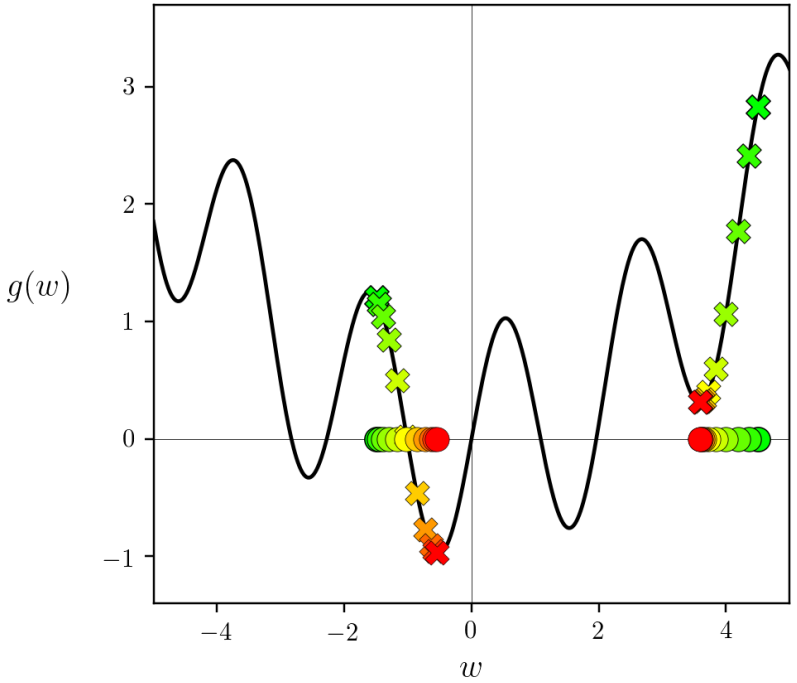
L'algoritmo ha due parametri: η (passo) e T (numero di passi) (chiamati *iperparametri* per non confonderli con i parametri w del modello da ottimizzare)

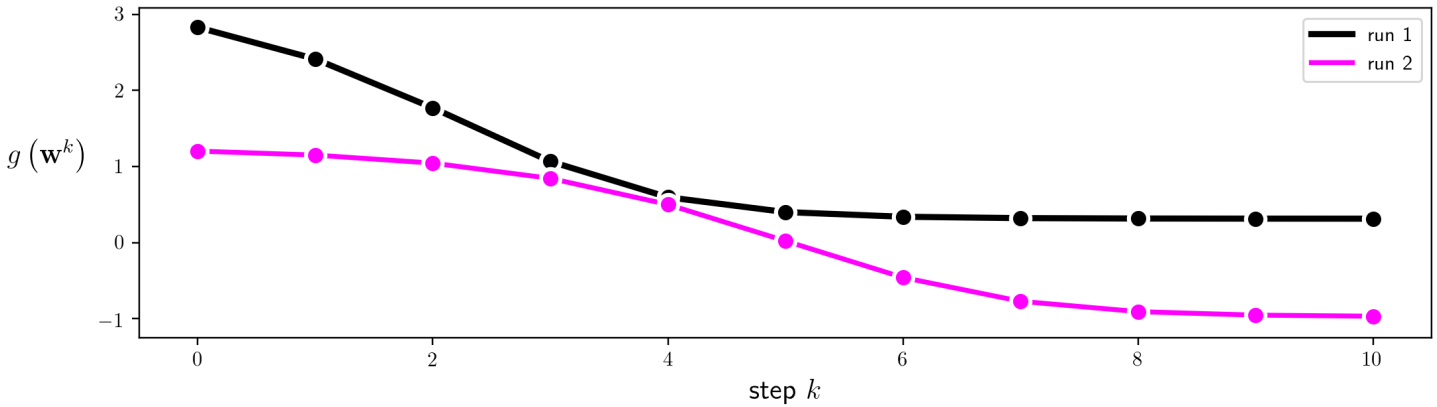
Esempio con diversi valori del passo η



Esempio non convesso





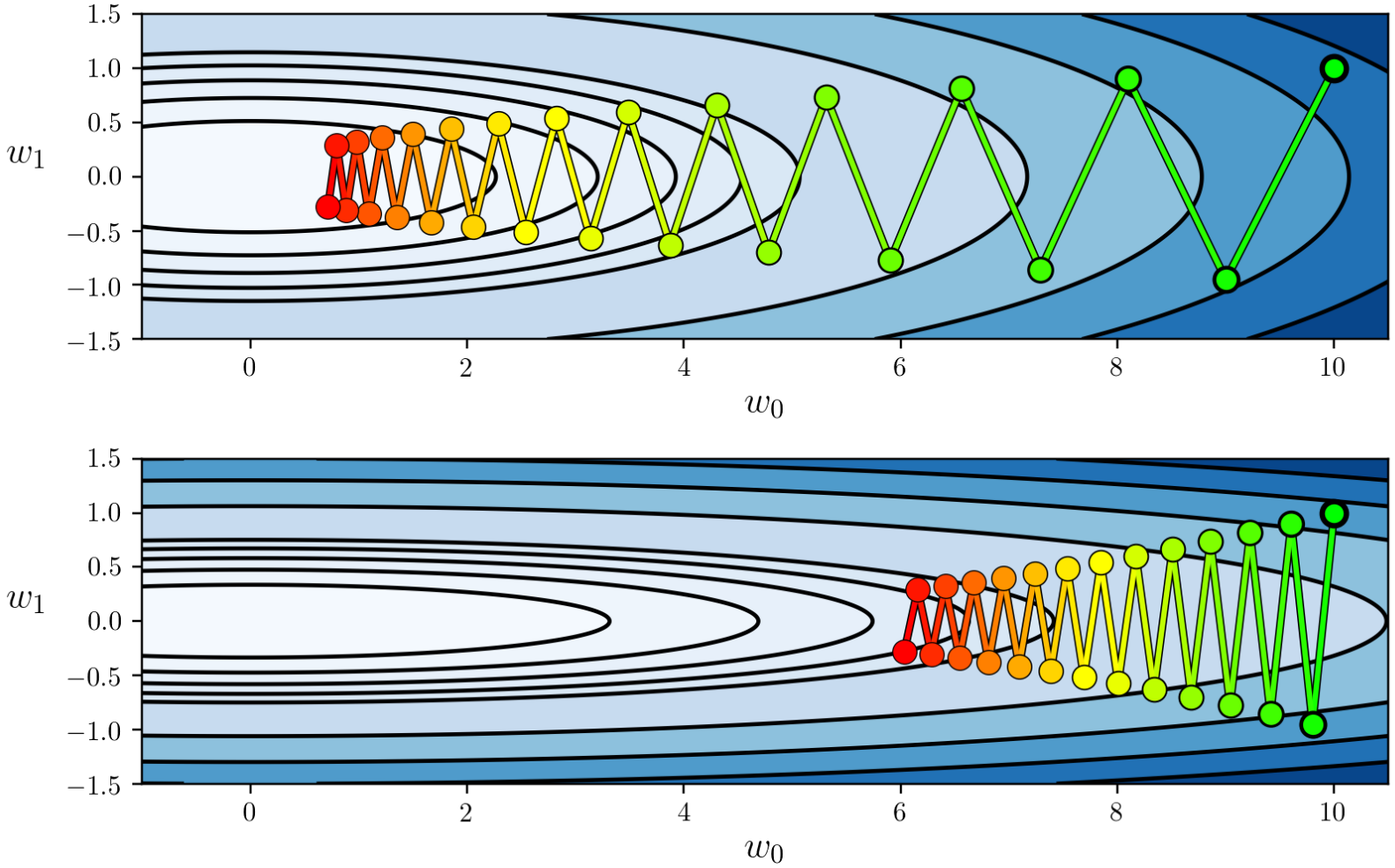


Due problematiche di GD

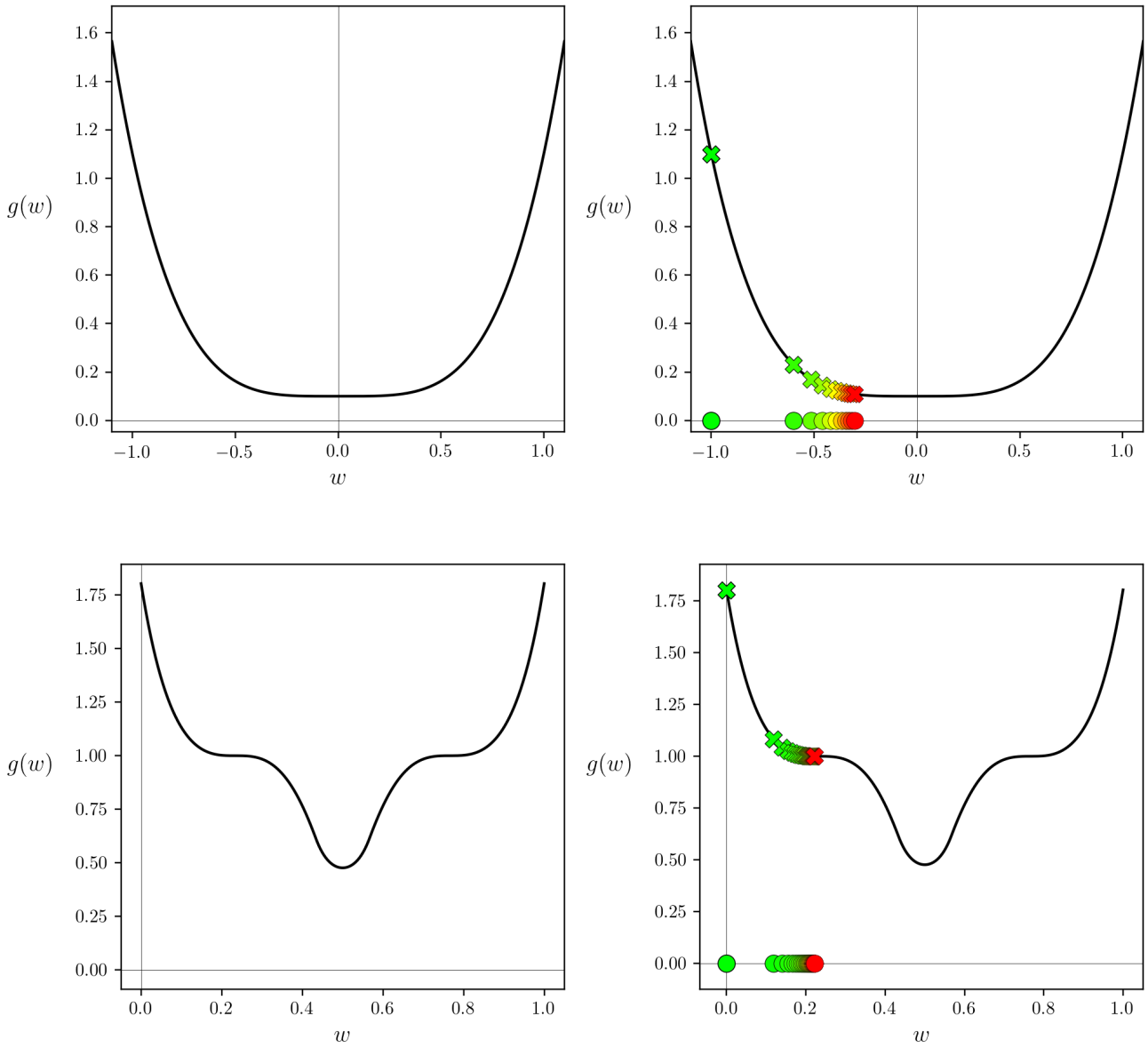
- La direzione del gradiente negativo può oscillare, portando l'algoritmo a muoversi a "zig-zag" e convergere lentamente
- La magnitudine del gradiente negativo si contrae vicino ai punti critici, rallentando la discesa

Per questo motivo le librerie di ML adottano varianti di GD più sofisticate (Momentum, Adam, RMSprop...)

Movimento a zig-zag: Esempi



Rallentamento vicino ai punti critici: Esempi



Metodi del secondo ordine

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

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

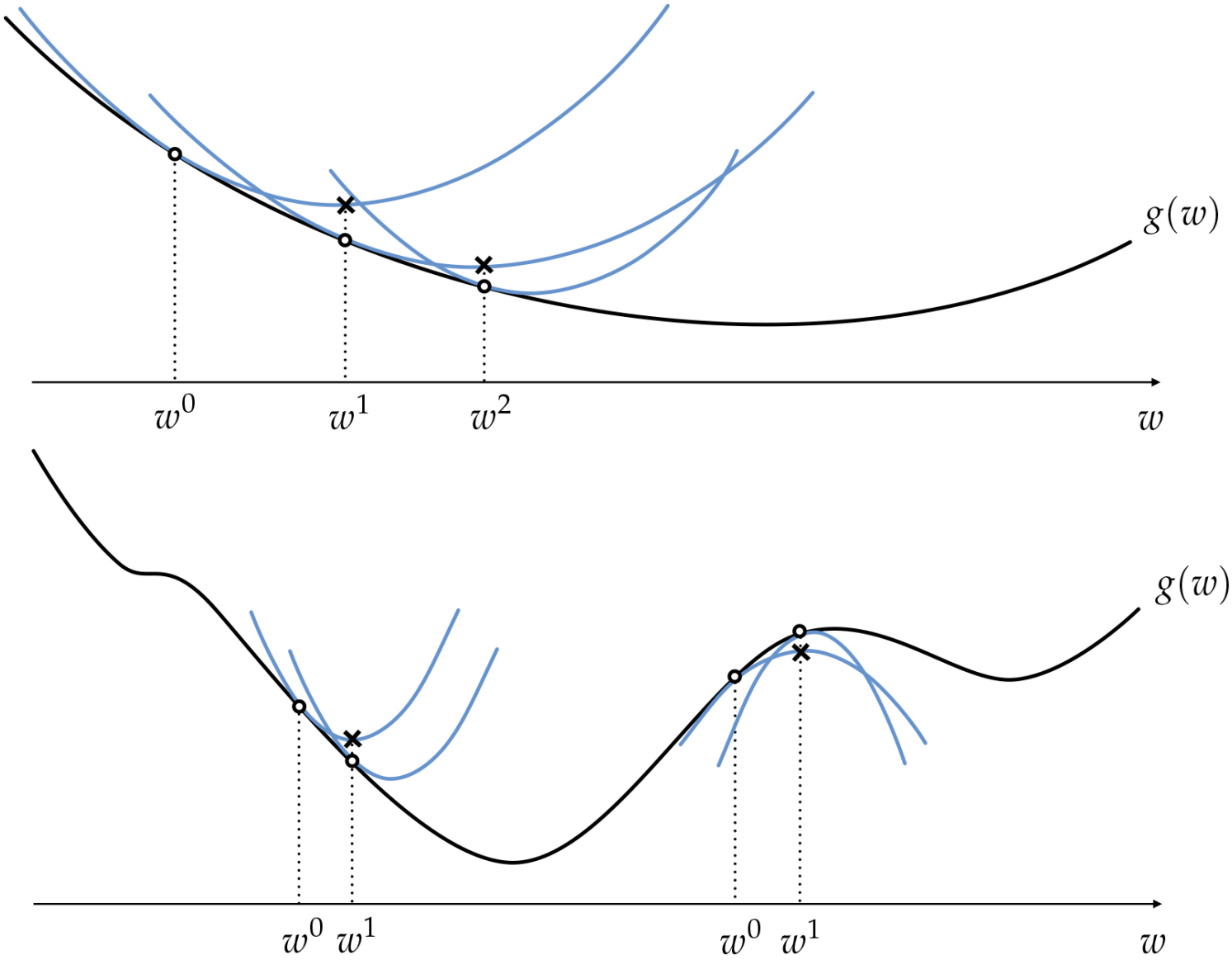
I *metodi del secondo ordine* utilizzano anche

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

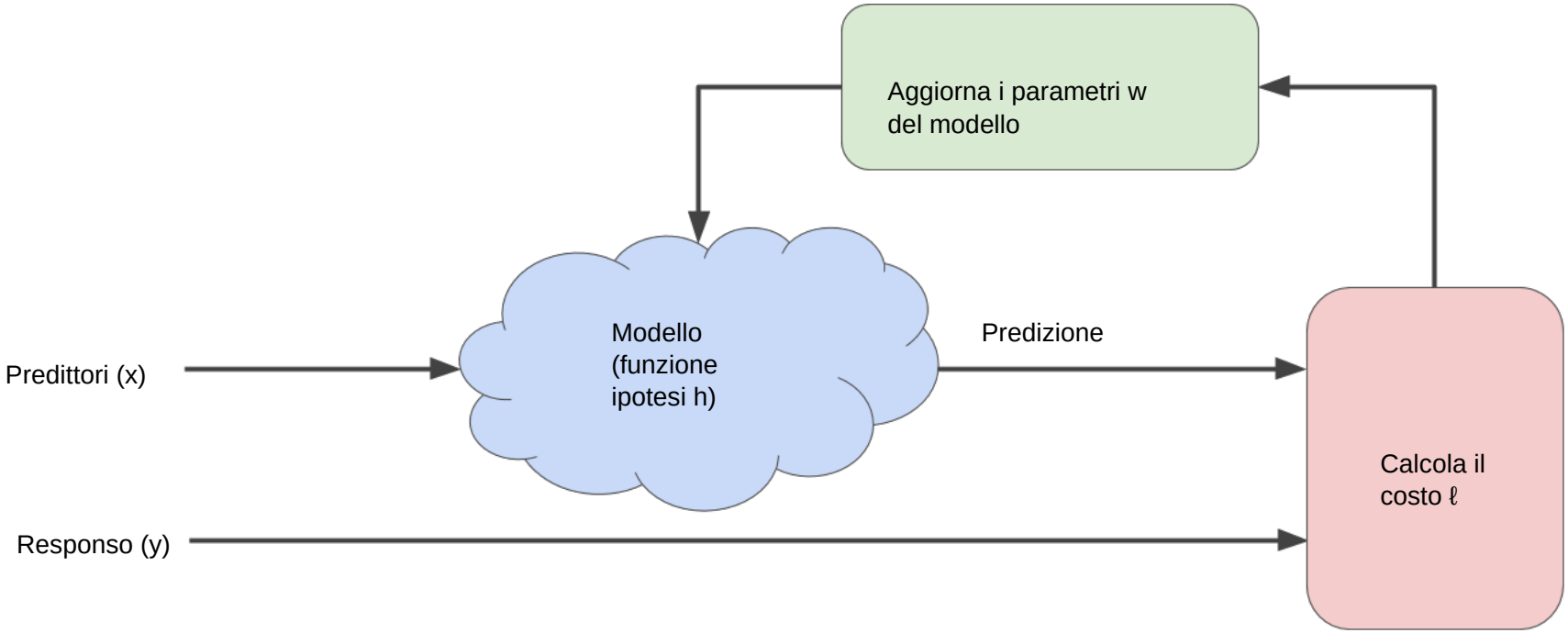
Il *metodo di Newton* ne è l'esempio più noto

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

Esempio (metodo di Newton)



Minimizzazione iterativa del rischio empirico



Discesa del gradiente (GD): calcolo di un passo

Il metodo GD calcola $\nabla g(w)$ ad ogni passo

Per noi, g è il rischio empirico, funzione di tutti gli esempi di training:

$$g(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(h_w, (x^{(i)}, y^{(i)})) = \frac{1}{m}\sum_{i=1}^{m} \ell_i(h_w)$$

$$\nabla g(w) = \frac{1}{m}\sum_{i=1}^{m} \nabla \ell_i(h_w)$$

dove h_w è l'ipotesi codificata dal vettore w e ℓ_i è la funzione di costo sull'esempio i-esimo

Discesa del gradiente (GD) per il Machine Learning

Gradient Descent (Batch GD)

1. Poni $w^{(1)} = 0$
2. Per $t = 1, \ldots, T$:
   - Poni $w^{(t+1)} = w^{(t)} - \eta \cdot \frac{1}{m} \sum_i \nabla \ell_i(h_{w^{(t)}})$
3. Restituisci il $w^{(t)}$ col minimo valore di $g(w^{(t)})$, $t = 1, \ldots, T$

Ogni passo del metodo GD richiede di considerare tutti gli m esempi (si parla di metodo *batch*)

Discesa stocastica del gradiente (SGD)

Per dei passi più rapidi, si usa una variante stocastica di GD

Stochastic Gradient Descent (SGD)

1. Poni $w^{(1)} = 0$
2. Per $t = 1, \ldots, T$:
   - Estrai un esempio i a caso
   - Poni $w^{(t+1)} = w^{(t)} - \eta \cdot \nabla \ell_i(h_{w^{(t)}})$
3. Restituisci il $w^{(t)}$ col minimo valore di $g(w^{(t)})$, $t = 1, \ldots, T$

Ogni passo del metodo SGD richiede di considerare un solo esempio

Grande risparmio computazionale rispetto al metodo GD batch; la localizzazione dell'ottimo diviene meno precisa

Mini-batch SGD

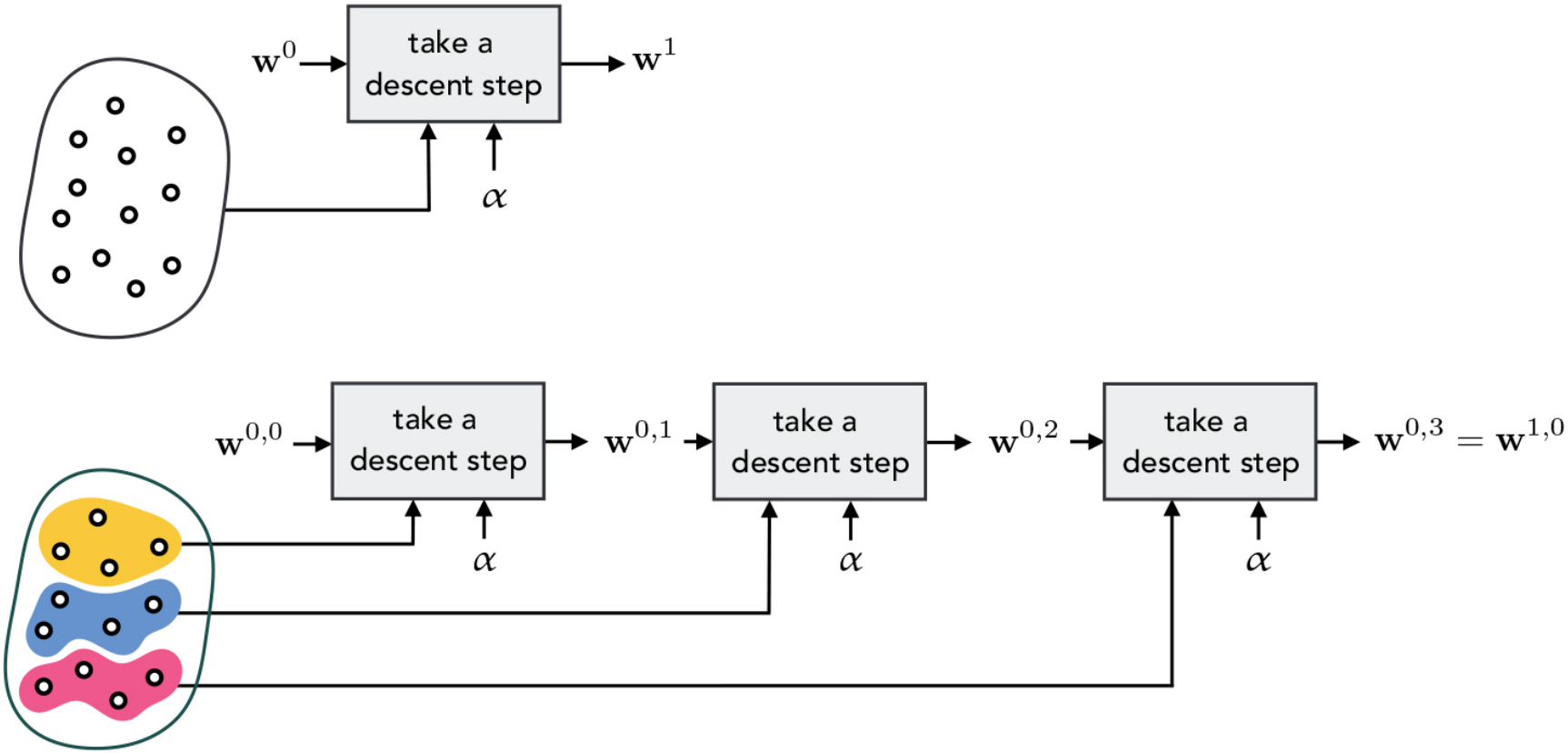
Mini-batch SGD è un compromesso tra GD e SGD

Mini-Batch SGD (per l'apprendimento supervisionato)

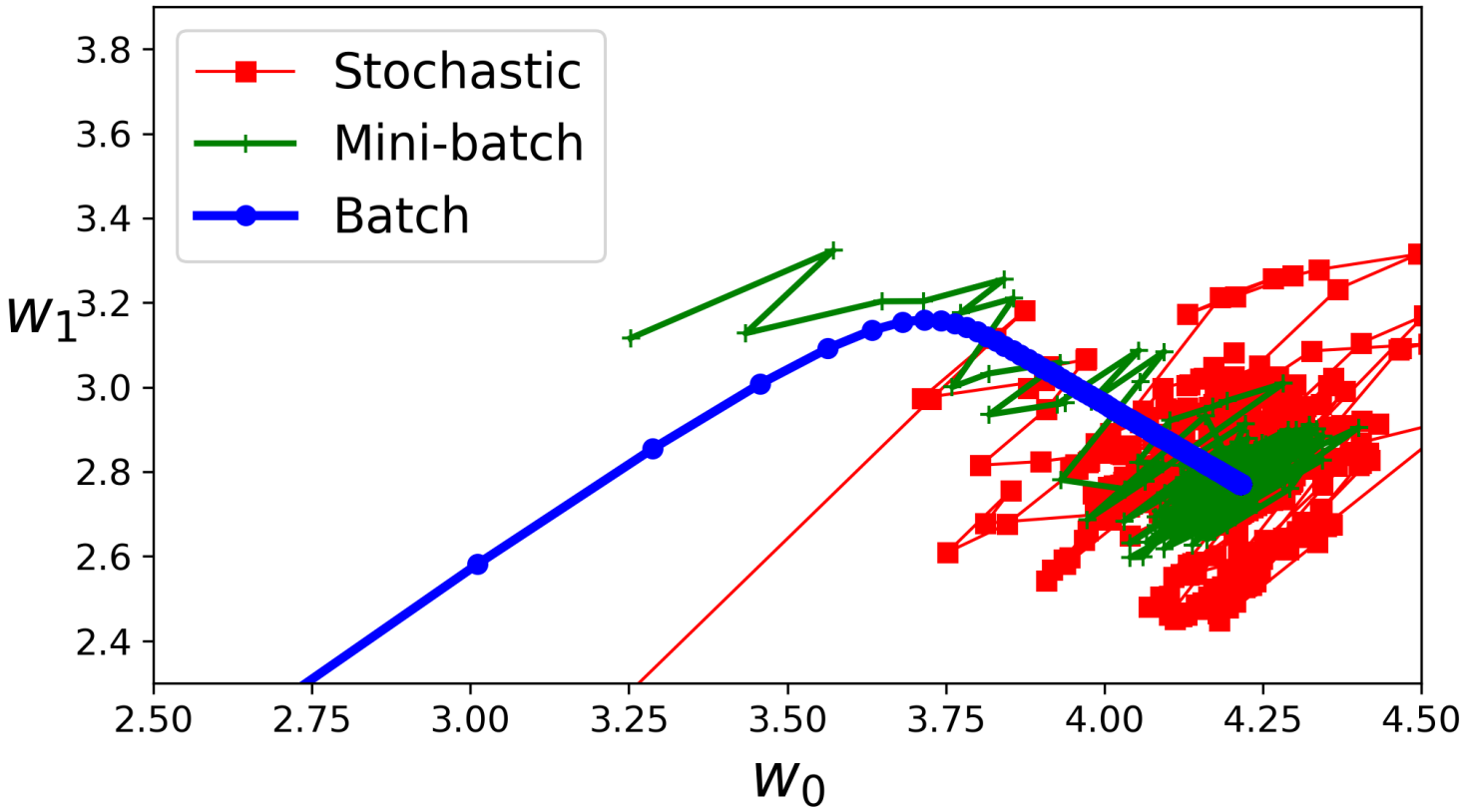
1. Poni $w^{(1)} = 0$
2. Per $t = 1, \ldots, T$:
   - Estrai un lotto (*batch*) B di esempi a caso
   - Poni $w^{(t+1)} = w^{(t)} - \eta \cdot \frac{1}{|B|} \sum_{i \in B} \nabla \ell_i(h_{w^{(t)}})$
3. Restituisci il $w^{(t)}$ col minimo valore di $g(w^{(t)})$, $t = 1, \ldots, T$

Ogni passo di mini-batch SGD richiede di considerare $|B|$ esempi

Batch GD vs. Mini-Batch SGD



Batch GD vs. SGD vs. Mini-Batch SGD



Norme

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

1. $\|x\| = 0$ se e solo se $x = 0$
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 L_2 (euclidea): $\|x\|_2 = (\sum_j |x_j|^2)^{1/2}$
- norma L_1: $\|x\|_1 = \sum_j |x_j|$
- norma L_p: $\|x\|_p = (\sum_j |x_j|^p)^{1/p} \qquad (p \geq 1)$
- norma L_∞ (uniforme): $\|x\|_\infty = \max_j |x_j|$

Ogni norma è una funzione convessa (perché?)