Discesa del gradiente per l'apprendimento supervisionato

Vincenzo Bonifaci

IN550 – Machine Learning

Regressione con altre funzioni di costo

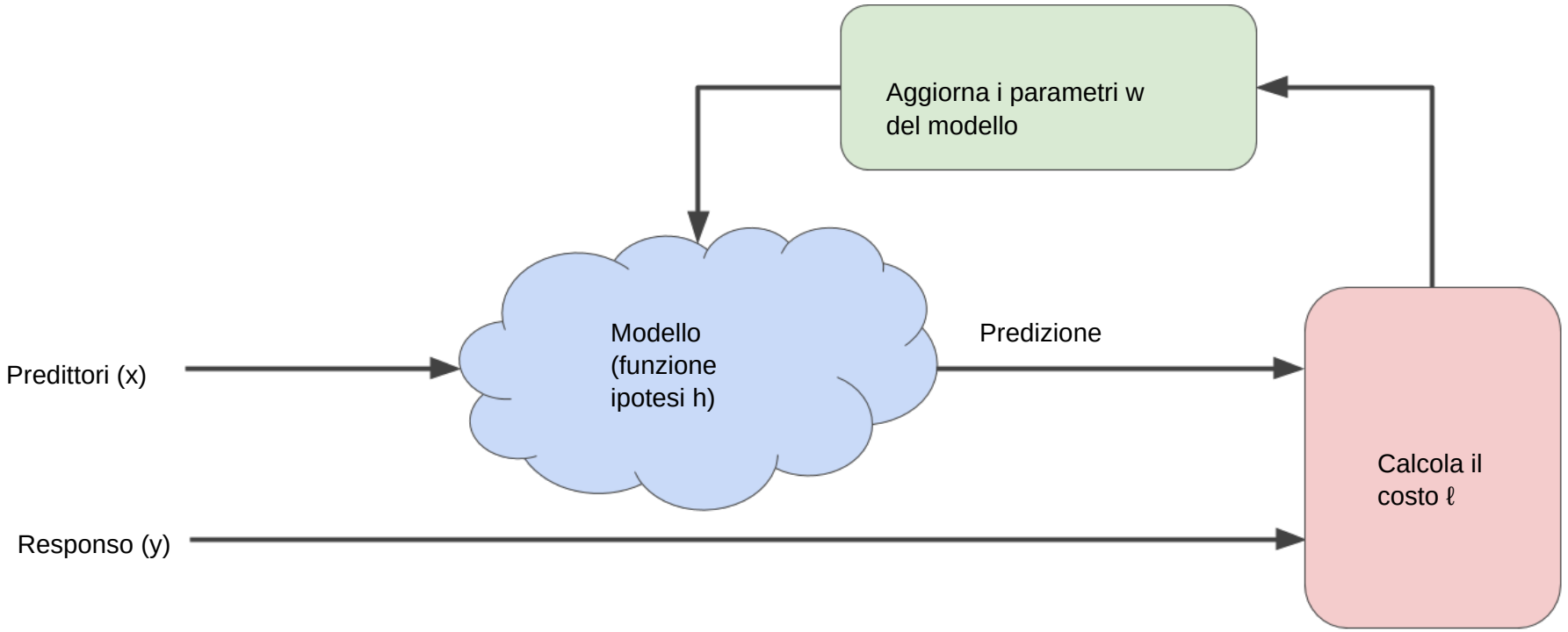
Come trattare funzioni di costo diverse da quella quadratica?

Per esempio, nella regressione *Least Absolute Deviation* (*LAD*),

$$\ell(h, (x, y)) \stackrel{\text{def}}{=} |h(x) - y|$$

Per una vasta classe di funzioni di costo (convesse e/o differenziabili) esiste una metodologia generale di ottimizzazione

Minimizzazione iterativa del rischio empirico



Discesa del gradiente (GD)

Sia $f : \mathbb{R}^d \to \mathbb{R}$ una funzione convessa differenziabile, con *gradiente*

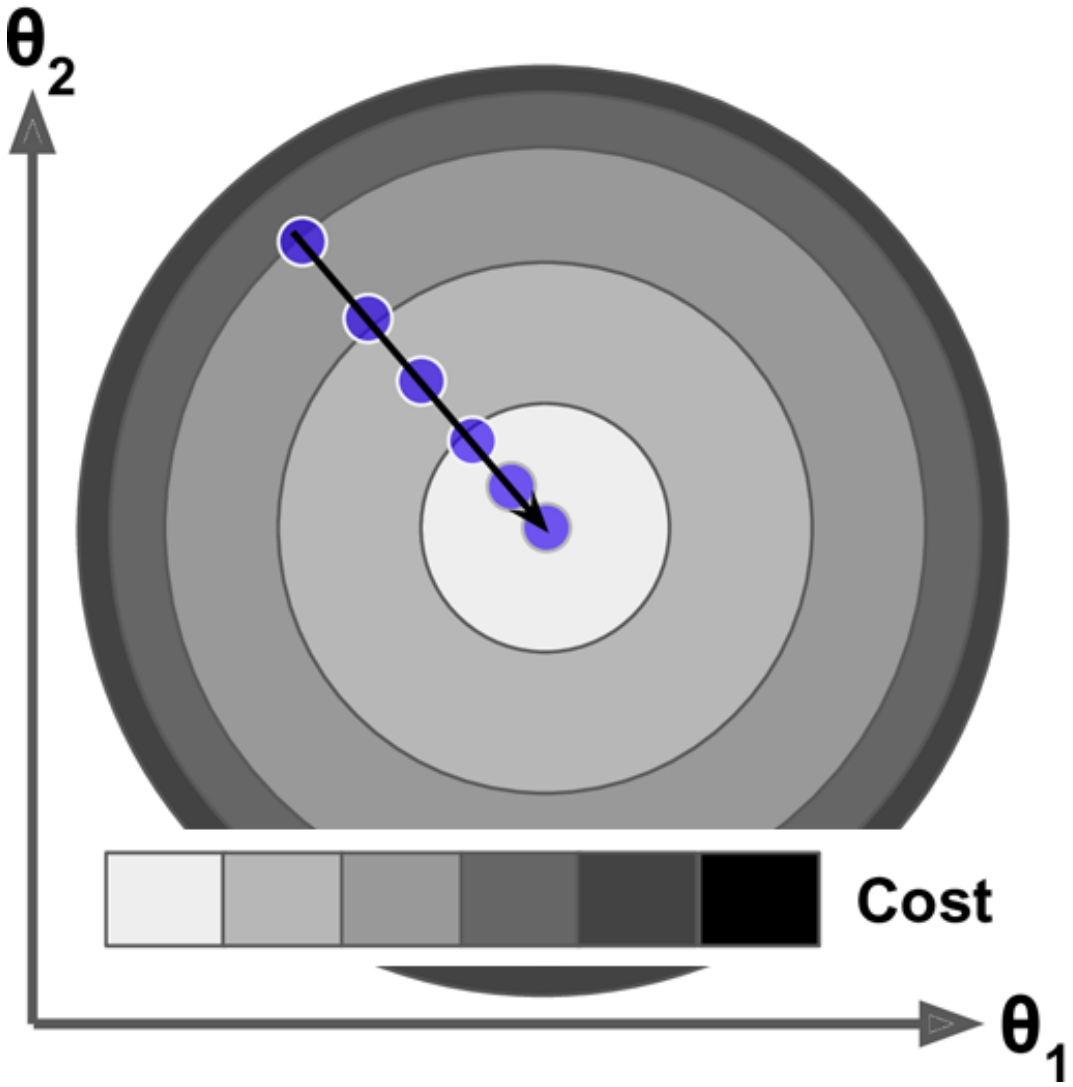
$$\nabla f(w) = \left(\frac{\partial f}{\partial w_1}(w), \ldots, \frac{\partial f}{\partial w_d}(w) \right)$$

Algoritmo Gradient Descent (generico)

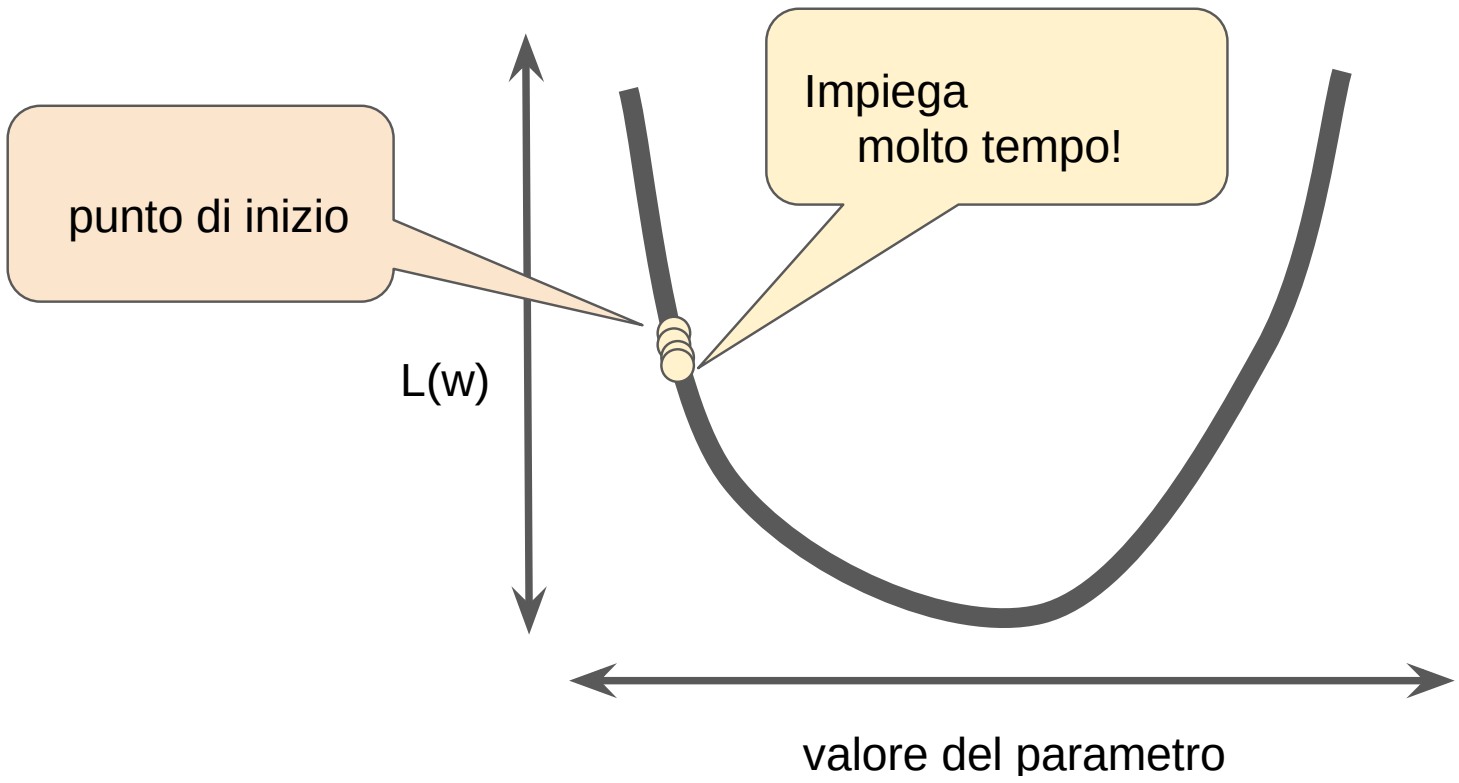
1. Poni $w^{(1)} = 0$
2. Per $t = 1, \ldots, T$: $w^{(t+1)} = w^{(t)} - \eta \cdot \nabla f(w^{(t)})$
3. Restituisci $w^{(T)}$

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

Discesa del gradiente (GD)

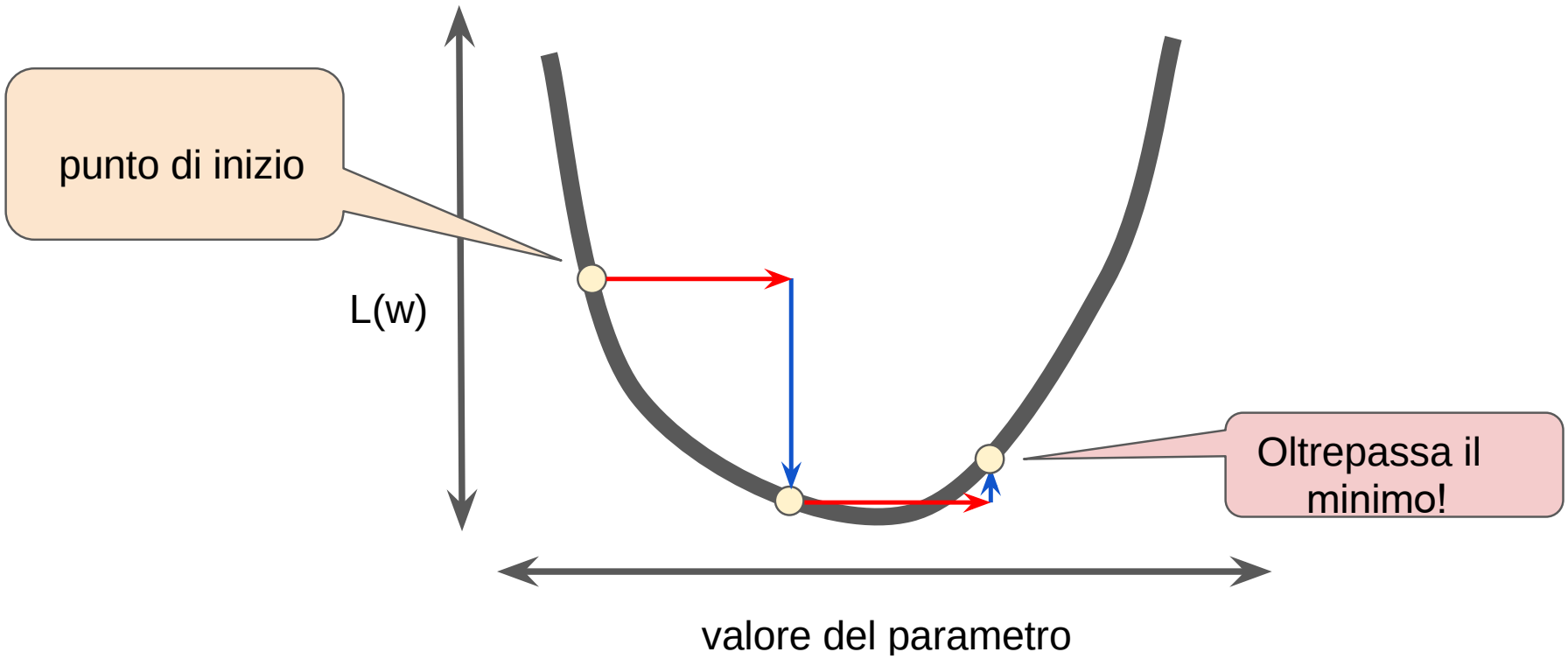


Discesa del gradiente (GD): impatto del tasso η



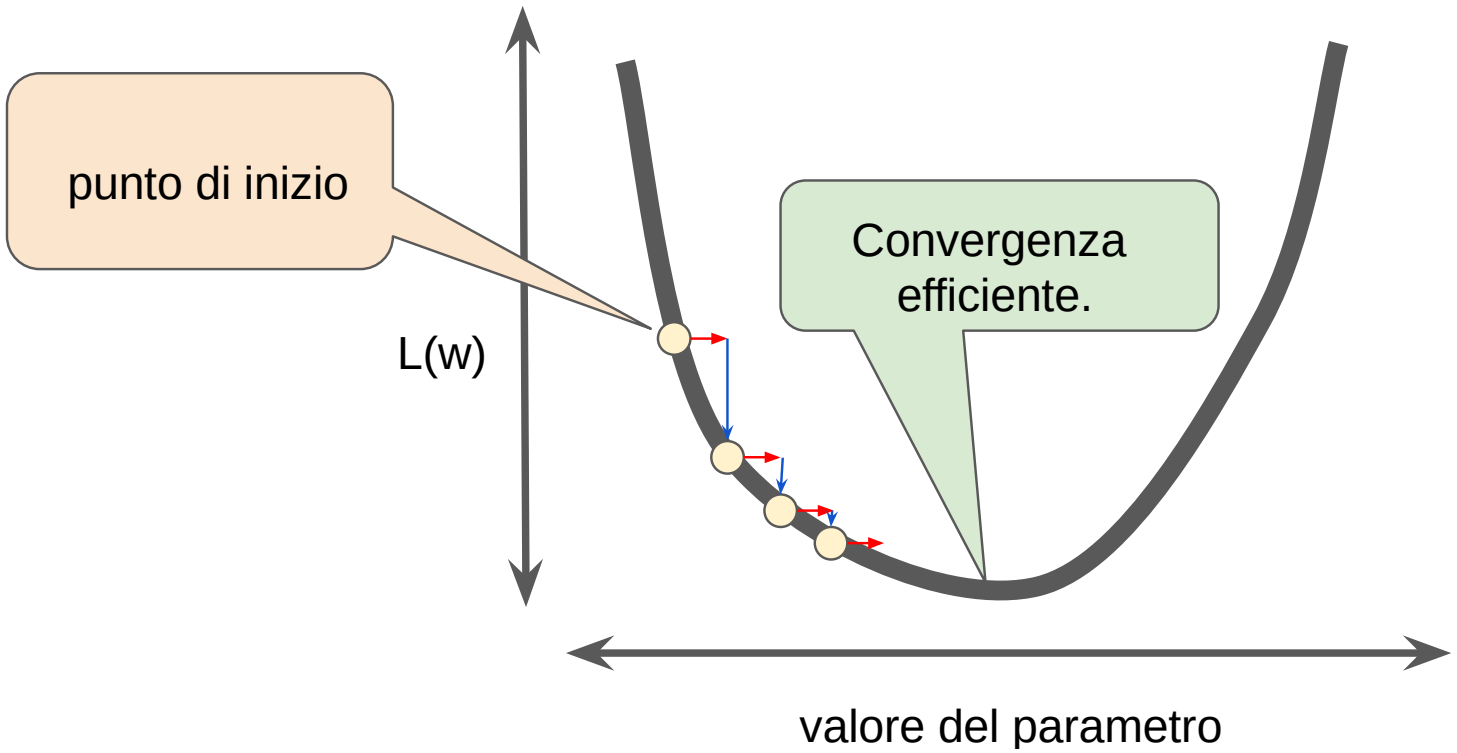
Tasso di apprendimento troppo basso

Discesa del gradiente (GD): impatto del tasso η



Tasso di apprendimento troppo alto

Discesa del gradiente (GD): impatto del tasso η



Tasso di apprendimento adeguato

Discesa del gradiente (GD): Calcolo di un passo

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

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

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

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

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

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

Discesa del gradiente (GD) riformulato

Gradient Descent (per l'apprendimento supervisionato)

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 $w^{(T)}$

Discesa stocastica del gradiente (SGD)

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

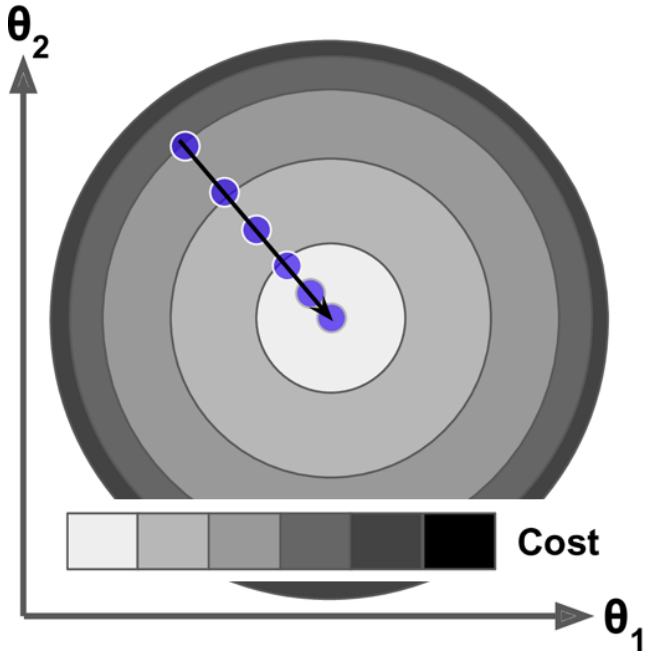
Stochastic Gradient Descent (per l'apprendimento supervisionato)

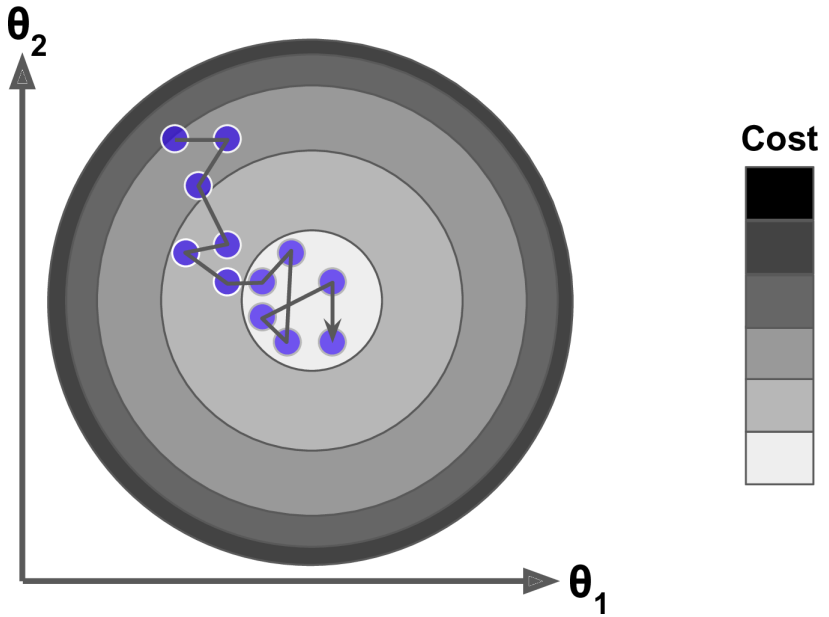
1. Poni $w^{(1)} = 0$
2. Per $t = 1, \ldots, T$:
   - Estrai un esempio i a caso
   - Poni $w^{(t+1)} = w^{(t)} - \eta \cdot \nabla \ell_i(h_{w^{(t)}})$
3. Restituisci $w^{(T)}$

Ogni passo del metodo SGD richiede di considerare un solo esempio

Poiché $\mathbb{E}_i[\nabla \ell_i(h)] = \frac{1}{m} \sum_{i=1}^{m} \nabla \ell_i(h)$, l'aggiornamento è in valore atteso lo stesso di GD

GD vs. SGD





Esempio: SGD per la regressione lineare

Esempio

Derivare la regola di aggiornamento di SGD nel seguente scenario:

- Ipotesi lineari: $h_w(x) = w^\top x$
- Costo quadratico: $\ell(h_w) = (h_w(x) - y)^2$

Calcolando il gradiente di ℓ, otteniamo la regola di aggiornamento

Regola Least-Mean-Squares (LMS)

$$w \leftarrow w - 2\eta \cdot (w^\top x - y) \cdot x$$

Esempio: SGD per la regressione LAD

Esempio

Derivare la regola di aggiornamento di SGD nel seguente scenario:

- Ipotesi lineari: $h_w(x) = w^\top x$
- Costo LAD: $\ell(h_w) = |h_w(x) - y|$

Calcolando il gradiente di ℓ, otteniamo la regola di aggiornamento

$$w \leftarrow w - \eta \cdot \operatorname{sgn}(w^\top x - y) \cdot x$$

$$\text{dove } (\operatorname{sgn}(z))_j = \begin{cases} +1 & \text{se } z_j > 0 \\ -1 & \text{se } z_j < 0 \\ 0 & \text{se } z_j = 0 \end{cases}$$

Mini-batch SGD

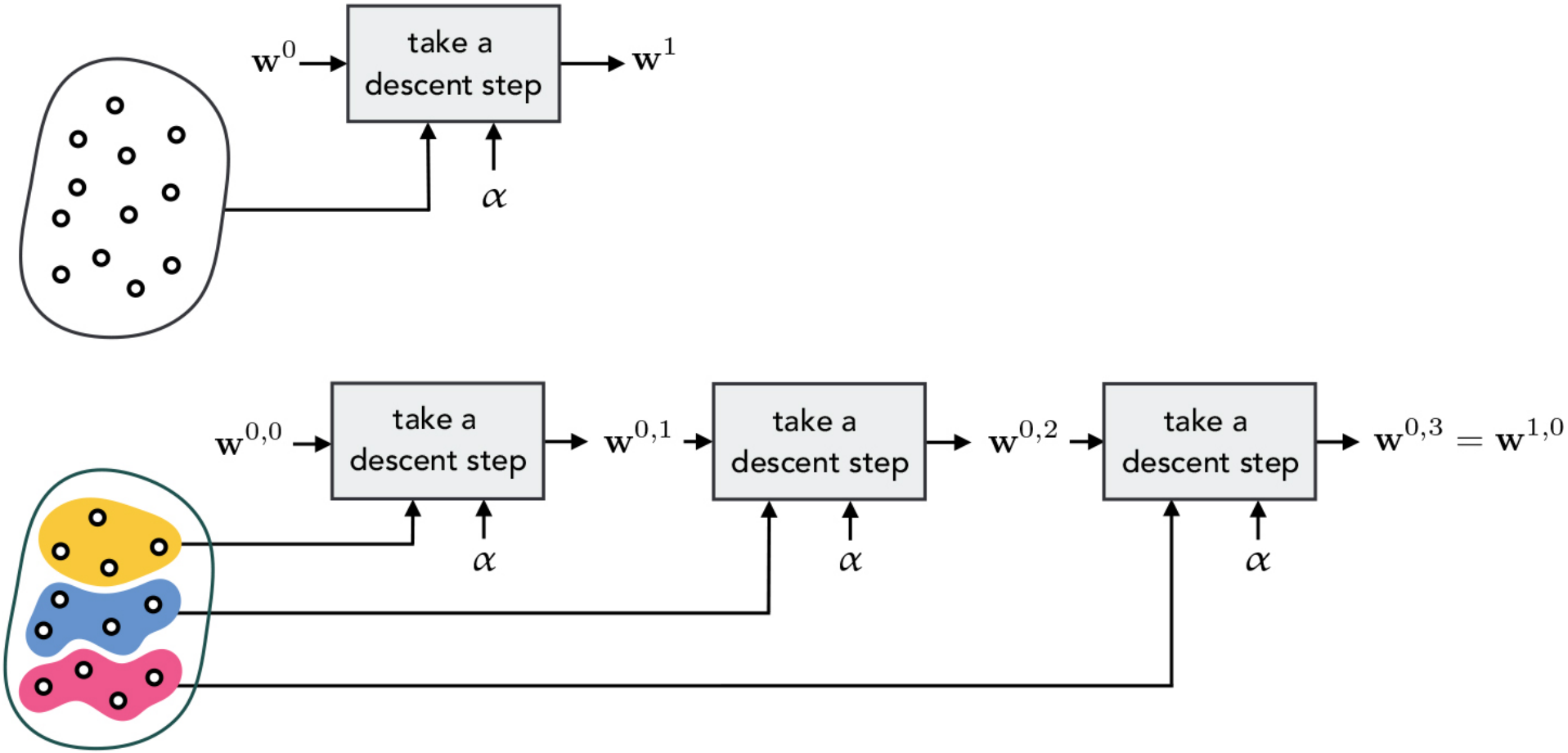
Mini-batch SGD è un compromesso tra GD e SGD

Mini-Batch SGD (per l'apprendimento supervisionato)

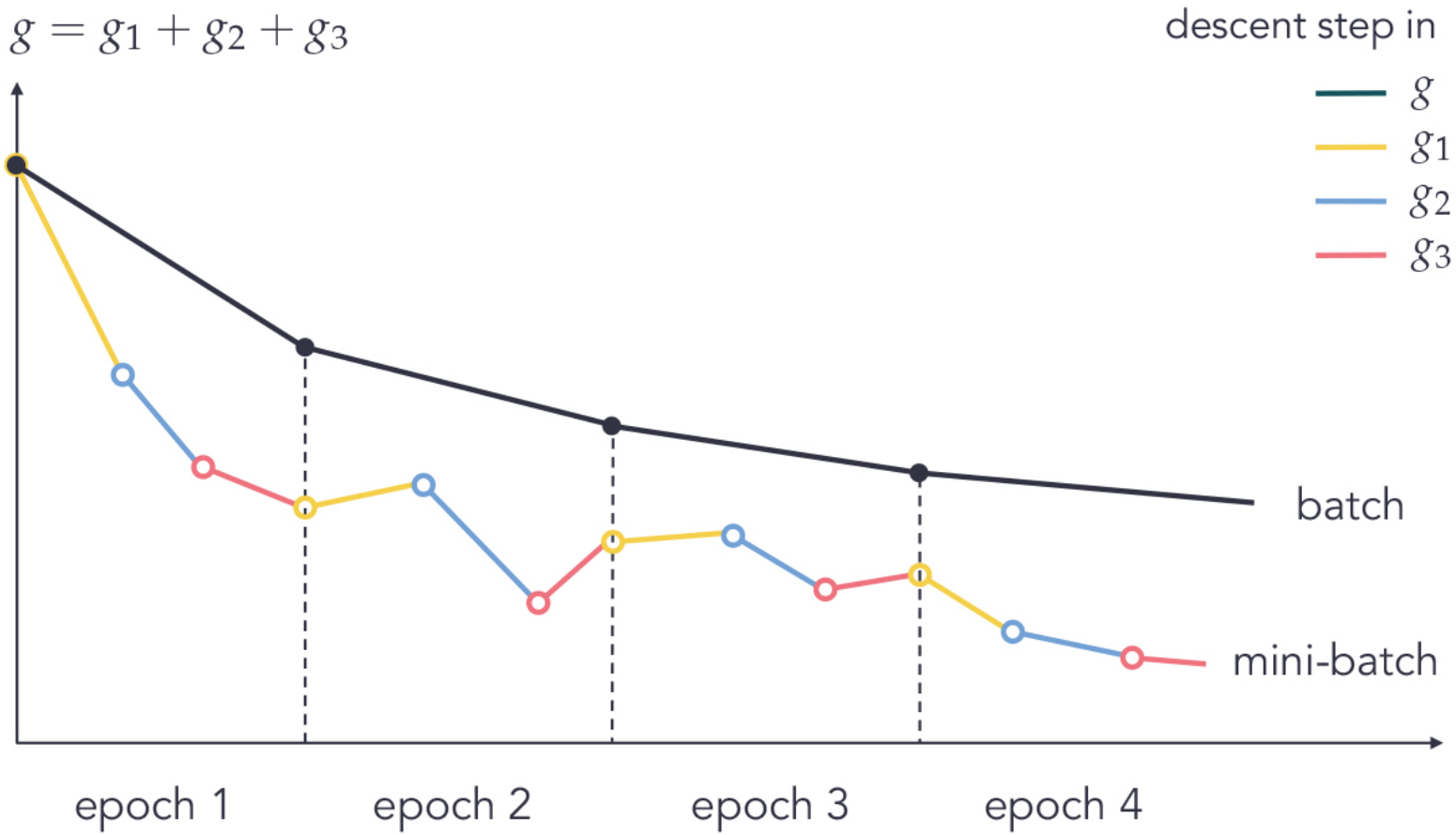
1. Poni $w^{(1)} = 0$
2. Per $t = 1, \ldots, T$:
   - Estrai un lotto (*batch*) B di esempi a caso
   - Poni $w^{(t+1)} = w^{(t)} - \eta \cdot \frac{1}{|B|} \sum_{i \in B} \nabla \ell_i(h_{w^{(t)}})$
3. Restituisci $w^{(T)}$

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

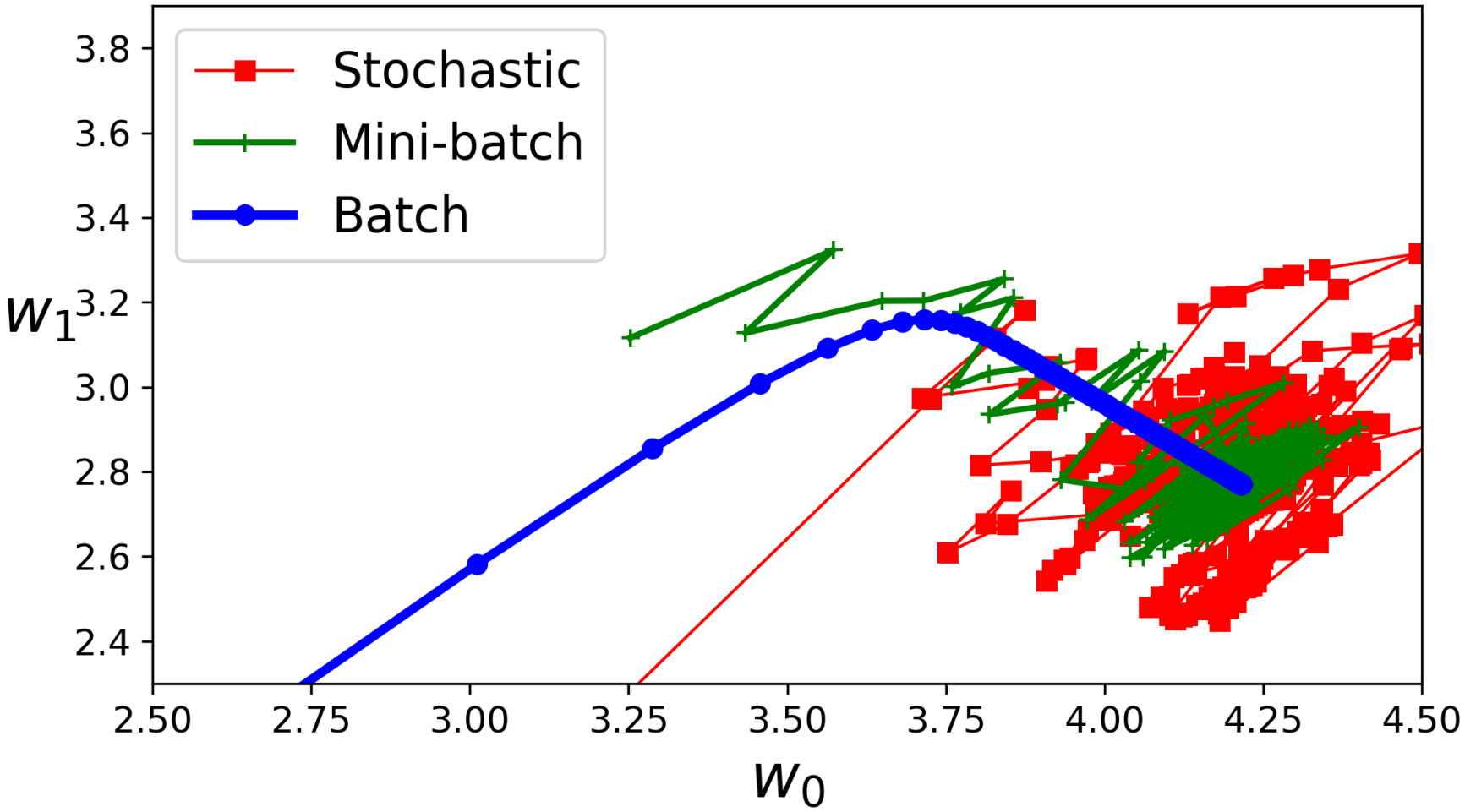
Batch GD vs. Mini-Batch SGD



Batch GD vs. Mini-Batch SGD



Batch GD vs. SGD vs. Mini-Batch SGD



Scalabilità computazionale dei metodi di regressione

m: numero di esempi
d: numero di variabili di input

Algoritmo	Esempi scanditi ad ogni passo	Dipendenza da d	Iperparametri	Interfaccia `scikit-learn`
Diretto (SVD)	m	$\Omega(d^2)$	nessuno	`LinearRegression`
Batch GD	m	$\Theta(d)$	η, T	`SGDRegressor`
SGD	1	$\Theta(d)$	η, T	`SGDRegressor`
Mini-batch SGD	$\|B\|$	$\Theta(d)$	$\eta, T, \|B\|$	`SGDRegressor`
K-NN	m	$\Theta(d)$	K	`KNeighborsRegressor`