

dim | (teo)

$$f(\bar{w}) - f(w^*) = f\left(\frac{1}{N} \sum_{n=1}^N w_n\right) - f(w^*)$$

$$(\text{Jensen}) \rightarrow \leq \frac{1}{N} \sum_{n=1}^N f(w_n) - f(w^*)$$

$$= \frac{1}{N} \sum_{n=1}^N \underbrace{\left( f(w_n) - f(w^*) \right)}_{\nearrow}$$

$$\underbrace{\langle \nabla f(w_n), w_n - w^* \rangle}_{\substack{\parallel \\ v_n}}$$

$$\left. \begin{array}{l} \text{lemma precedente} \\ \text{e } \parallel \nabla f(w_n) \parallel \leq \rho \\ \text{f è } \rho\text{-Lipsch.} \end{array} \right\} \leq \frac{R\rho}{\sqrt{N}}$$

definizione | Data  $f$  tale che

$$f(u) \geq f(w) + \langle \nabla f(w), u - w \rangle \quad \forall u$$

per un certo  $v$ , allora tale  $v$  si dice Sottogradiente di  $f$  in  $w$ .

Denotiamo con  $\partial f(w)$  l'insieme dei

sottogradienti  $v$  di  $f$  in  $w$ .

Oss. |  $f: A \rightarrow \mathbb{R}$  convessa, con  $A$  aperto e convesso

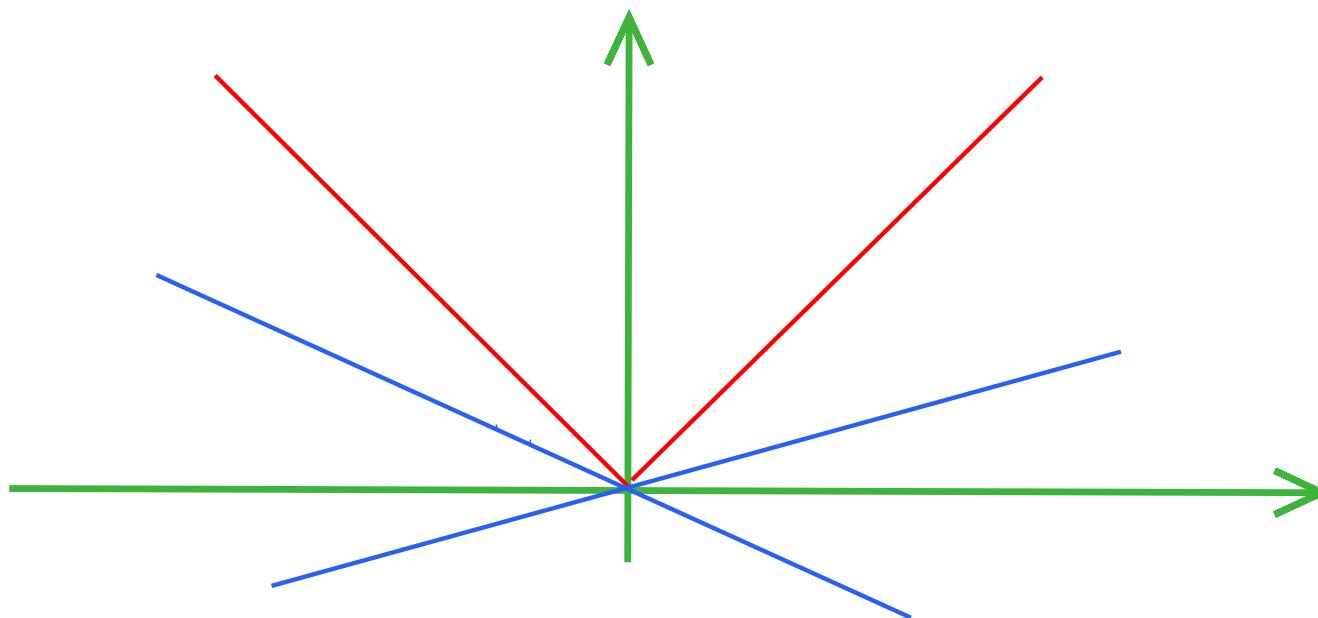


$$\partial f(w) \neq \emptyset \quad \forall w \in A$$

Inoltre, se  $f$  è differenziabile,  
 $\nabla f(w)$  è l'unico elemento di  $\partial f(w)$ .

es. |  $f(x) = |x|$

$$\partial f(x) = \begin{cases} \{1\} & \text{se } x > 0 \\ \{-1\} & \text{se } x < 0 \\ [-1, 1] & \text{se } x = 0 \end{cases}$$



$$\text{IDEA : } w_{n+1} = w_n - \eta \cdot v_n \quad (\Delta) \quad \text{Lo} \in \partial f(w_n)$$

lemma | Sia  $A \subset \mathbb{R}^d$  aperto e convesso.

Sia  $f: A \rightarrow \mathbb{R}$  convessa e  $\rho$ -Lipschitziana.

Allora  $\forall w \in A$  vale

$$\|v\| \leq \rho \quad \forall v \in \partial f(w)$$

dim | fissiamo  $w \in A$ ,  $v \in \partial f(w)$ .

$A$  aperto  $\Rightarrow \exists \varepsilon > 0$  t.c.

$$u := w + \varepsilon \frac{v}{\|v\|} \in A$$

inoltre

$$\langle u - w, v \rangle = \varepsilon \|v\|$$

$$\|v - w\| = \varepsilon$$

$$\cancel{\rho\varepsilon} = \rho \|u - w\| \geq f(u) - f(w) \geq \langle v, u - w \rangle = \varepsilon \|v\|$$

$\uparrow$

$v \in \partial f(w) \#$

! Vale il teorema precedente anche

per  $f: A \rightarrow \mathbb{R}$  convessa e  $\rho$ -Lipschitz.

Prendendo  $w_n$  come in  $(\Delta)$

• Versione stocastica:

Idea:  $v_n \in \partial f(w_n) \rightsquigarrow V_n$  il cui v.a. condiz.  
v.a      |  
              sta in  $\partial f(w_n)$

- Costruzione:

•  $V = (V_n)_{n=1, \dots, N-1}$  processo stocastico

su  $(\Omega, \mathcal{F}, \mathbb{P})$

• Poniamo:

$$W_1 = 0$$

$$W_{n+1} := W_n - \underbrace{\eta}_{V_n} \underbrace{\downarrow}_0 , \quad n = 1, \dots, N-1$$

• definiamo:

$$\mathcal{F}_n := \mathcal{F}_n^W , \quad n = 1, \dots, N$$

filtrazione naturale di  $W$

!  $(V_n)_n$  non è adattato a  $(\mathcal{F}_n)_n$

H.  $\mathbb{E}[V_n | \mathcal{F}_n] \in \underbrace{\partial f(w_n)}_{\text{insieme aleatorio}} \quad \text{P-q.c.}$

$\forall n = 1, \dots, N-1$

↓  
insieme aleatorio

V.a

DISCESA DET.:  $w_{n+1} = w_n - \gamma v_n \leftarrow$  vettori deturm.  
 $\partial f(w_n)$

DISCESA STOC.:  $w_{n+1} = w_n - \gamma v_n \leftarrow$  vettori aleatori  
 $\partial f(w_n)$   
 $\subset$  insieme aleatorio

$$v_n(\omega) \notin \partial f(w_n(\omega))$$

$$\mathbb{E}[v_n | \mathcal{F}_n] \in \partial f(w_n)$$

esempio Definiamo  $v_n := g(w_n, z_n)$  dove:

- $(z_n)_{n=1, \dots, N-1}$  v.a i.i.d.
- $g: A \times \mathbb{R}^q \rightarrow \mathbb{R}^d$  misurabile e  
 $\mathbb{R}^d$  tale che

$$\mathbb{E} [g(w, z_1)] \in \partial f(w) \quad \forall w \in A$$

↓

$W = (w_n)_{n=1,\dots,N}$  ha la prop.

di Markov

!

$$\mathbb{E}[V_n | \mathcal{G}_n] = \mathbb{E}[g(w_n, z_n) | \mathcal{G}_n]$$

" indip.  
 $y_w$   $\cap$   
 $\mathcal{G}_n$  ~~indip.~~

$$\left( \begin{array}{c} \text{freezing} \\ \text{lemma} \end{array} \right) \rightarrow = \mathbb{E}[g(w, z_n)]|_{w=w_n}$$

$$\cap \quad \nearrow \quad (\text{H})$$

$$\partial f(w_n)$$

teorema | Sia  $A$  aperto e convesso.

Sia  $f: A \rightarrow \mathbb{R}$  convessa e  
p-Lipschitziana. Sia

$w^* \in A$  t.c.  $\|w^*\| \leq R$  con  $R > 0$ .

Assumendo che  $\eta := \frac{R}{\rho \sqrt{N}}$  e

che  $\|V_n\| \leq \rho$  p-q.c.  $\forall n$ .

Allora, sotto H. vale che

$$\mathbb{E} \left[ f(\bar{w}) \right] - f(w^*) \leq \frac{R\rho}{\sqrt{N}}$$

||

$$\frac{1}{N} \sum_{n=1}^N w_n$$

corollario] Sotto le ipotesi del teorema precedente,

$$w^* \in \arg \min_{\|w\| \leq R} f(w) \Rightarrow \left| \mathbb{E}[f(\bar{w})] - f(w^*) \right| \leq \frac{R\rho}{\sqrt{N}}$$

dipende da N

In particolare:

$$N \geq R^2 \rho^2 / \varepsilon^2 \Rightarrow \left| \mathbb{E}[f(\bar{w})] - f(w^*) \right| \leq \varepsilon$$

dim [teo)

$$\mathbb{E} \left[ f(\bar{w}) - f(w^*) \right] \leq \mathbb{E} \left[ \underbrace{\frac{1}{N} \sum_{n=1}^N (f(w_n) - f(w^*))}_{\textcircled{11} \text{ & Jensen}} \right]$$

$$\frac{1}{N} \sum_{n=1}^N (f(w_n) - f(w^*))$$

$$= \frac{1}{N} \sum_{n=1}^N \underbrace{\mathbb{E} [f(w_n) - f(w^*)]}_{\text{II P-q.c.}}$$

$$\langle w_n - w^*, \underbrace{\mathbb{E}[v_n | \mathcal{F}_n]}_{\partial f(w_n)} \rangle$$

$$\stackrel{n \rightarrow (\text{H.})}{\partial f(w_n)}$$

$$\leq \frac{1}{N} \sum_{n=1}^N \underbrace{\mathbb{E} [\langle w_n - w^*, \mathbb{E}[v_n | \mathcal{F}_n] \rangle]}_{\text{II}}$$

proprietà  
torre

$$\mathbb{E} [\langle w_n - w^*, v_n > | \mathcal{F}_n ]$$

$$\hat{=} \frac{1}{N} \sum_{n=1}^N \mathbb{E} [\langle w_n - w^*, v_n >]$$

$$= \mathbb{E} \left[ \underbrace{\frac{1}{N} \sum_{n=1}^N \langle w_n - w^*, v_n >}_{\frac{R\rho}{\sqrt{N}}} \right]$$

II (lemma algebrico)  
 $\frac{R\rho}{\sqrt{N}}$  (con  $\|v_n\| \leq \rho$ )

$$\leq \frac{R\rho}{\sqrt{N}} \quad \#$$

## • Applicazione al machine learning:

- Input: - un insieme dominio  $X$ 
  - " " " di caratteristiche  $Y$
  - un "training set"  $S$ .  
 $S = ((x_n, y_n))_{n=1, \dots, m}$   
 $\uparrow$   
 $X \times Y$
- Output:  $h: X \rightarrow Y$ 
  - ↳ classificatore
- L'insieme delle  $h$  possibili è  $H$
- Modello di dati:  $P$  misura di probabilità su  $X$  e  $f: X \rightarrow Y$ 
  - ↓
  - corretta classificazione
- Misure di successo: quanto è "corretta"  $h$   
es:  $L_{P,f}(h) := P(h \neq f)$

OSS.

GENERALIZZAZIONE:

$f \not\exists$ ,  $P$  probabilità su  $X \times Y$

!  $P$  non è nota al learner

- Altra misura di successo: rischio empirico

$$L_{S,f}(h) : \frac{1}{m} |\{i \in \{1, \dots, m\} : h(x_i) \neq f(x_i)\}|$$

$y_i$

- In generale:

$l : H \times \mathcal{Z} \rightarrow \mathbb{R}_+$  funzione costo  
 $(X \times Y, P)$

$$L_P(h) := \mathbb{E} \left[ l(h, \cdot) \right] \quad (P \text{ non è nota})$$

v.a.

es:  $\ell(h, (x, y)) := \begin{cases} 0 & \text{se } h(x) = y \\ 1 & \text{se } h(x) \neq y \end{cases}$

$$\ell(h, (x, y)) := (h(x) - y)^2$$

Problema:  $\min_{h \in H} L_P(h) \quad (*)$

- ASSOCIAZIONE:

$$\begin{array}{ccc} h & \longleftrightarrow & w \in \mathbb{R}^d \\ \cap & & \cap \\ H & \longleftrightarrow & A \subset \mathbb{R}^d \end{array}$$

$$(*) \quad \min_{w \in A} L_P(w)$$

definiz. Un problema  $(H, Z, l)$  si dice convesso se  $H$  è convesso e se  $l(\cdot, z)$  è convessa  $\forall z \in Z$ .

- Se  $(H, Z, l)$  è convesso allora  $L_P(\cdot)$  è convessa

$$(L_P(h) \rightarrow L_P(w) = \underset{\substack{\cap \\ \mathbb{R}^d}}{\mathbb{E}} [l(w, \cdot)])$$

- Discesa gradiente deterministica: non conosciamo  $P$ .
- Discesa stocastica:

$$V_n := \nabla l(W_n, Z_n), \text{ dove } Z = X \times Y$$

$(z_n)_n$  succ. di v.a. i.i.d t.c.

$z_n \sim p$

$$\mathbb{E}[v_n | \mathcal{F}_n] = \mathbb{E}[\nabla l(w_n, z_n) | \mathcal{F}_n]$$

$$(\text{freezing}) = \mathbb{E}[\nabla l(w, z_n)]|_{w=w_n}$$

$$! = \nabla \underbrace{\mathbb{E}[\ell(w, z_n)]}_{\text{II}}|_{w=w_n}$$

$$L_p(w_n)$$



H.