

## INTRO CLASSIFICAZIONE BINARIA

SI HANNO CLASSIFICATORI UNICI CHE WORKANO SU UN DATASET DEL TIPO:

$$D = \left\{ (x^{(i)}, y^{(i)}) \mid x^{(i)} \in \mathbb{R}^n, y^{(i)} \in \{-1, 1\}, i=1, \dots, N \right\}$$

IL DATASET È PUNTO DEFINITO DA UNA FUNKTION KNOWN  $f: \mathbb{R}^n \rightarrow \{-1, 1\}$  G  
TALCHÉ CHE  $f(x^{(i)}) = y^{(i)}$ . CERCHIAMO PUNTI CHE RISPONDENTI TALI CHE:

- $\text{sign}(h(x, w)) = f(x) \forall x$

- $\bar{w} = \underset{w}{\operatorname{argmax}} \mathcal{L}(w, D) + \lambda \mathcal{R}(w)$

LOSS

REGULARIZZAZIONE

NEL CASO DEI CLASSIFICATORI UNICI SI RICERCA PUNTI UN IPOPARABOLA CHE SODDISFA IL DATASET IN DUE PARTI SEPARATE IN MODO CHE SIANO CLASSIFICATORI CORRETTAMENTE (PER ESEMPIO).

DEFINISMO LA LOSS COME:

$$\mathcal{L}(w, D) = \frac{1}{N} \sum_{i=1}^N l(w^T x^{(i)} + b, y^{(i)})$$

DONDE  $l$  PUÒ ESSERE:

LOG-LOSS:  $\log(1 + e^{-y^{(i)}(w^T x^{(i)} + b)})$

→ CONTINUA MA NON DIFFERENZIABILE!

Hinge-loss

$$\max\{0, 1 - (w^T x^{(i)} + b)\}$$

Dove il segno di  $w^T x^{(i)} + b$  coincide con  $y^{(i)}$  se la classificazione è corretta.

SUM

Si ha il seguente problema:

$$\min_{w, b} \frac{1}{2} \|w\|^2 + \sum_{i=1}^N \max\{0, 1 - y^{(i)}(w^T x^{(i)} + b)\}$$

Dove  $y^{(i)}(w^T x^{(i)} + b)$  è una misura della Bontà della classificazione, e posto  
prodotto è negativo se le classi non coincidono, positivo altrimenti.

Siccome la funzione  $\max\{\cdot, 0\}$  non è differenziabile, si risolve il problema:

$$\min_{w, b, \varepsilon_i} \frac{1}{2} \|w\|^2 + C \sum_{i=0}^N \varepsilon_i$$

$$\cdot \varepsilon_i \geq 0 \quad \forall i = 1, \dots, N$$

$$\cdot y^{(i)}(w^T x^{(i)} + b) \geq 1 - \varepsilon_i \quad \forall i = 1, \dots, N$$

Dove i due problemi sono equivalenti, ovvero se  $(w^*, b^*)$  sono ottimali per il problema iniziale, allora  $(w^*, b^*, \varepsilon^*)$  è ottimale per il problema risolto. Infatti possiamo risolvere:

$$\varepsilon_i^* \geq \max\{0, 1 - y^{(i)}(w^T x^{(i)} + b)\}$$

INFATTI SE ABBANDONI  $\epsilon^*$ : raggiungere del massimo, potremmo non trovare un  $\epsilon$  esistente vicino al massimo e questo risultato sarebbe ammesso e con le funzioni obiettivo minore di quanto raggiunto prima con  $\epsilon^*$  ma questo è assurdo!

Quindi il problema risolto ci garantisce che la funzione obiettivo sia:

- **PENSATIVA**
- **CONVESA**
- **CONTINUA UNIPI**

SE consideriamo il caso in cui  $C = +\infty$  allora si dovrà avere  $\epsilon_i = 0$  per al fine di minimizzare, con  $y^{(i)}(w^T x^{(i)} + b) \geq 1$ , ovvero dovendo che tutti i

punti siano correttamente classificati e ben distanti dal margine.

Può essere tuttavia possibile avere so overfitting e inoltre il problema potrebbe non convergere soluzioni se le due classi non sono linearmente separabili.

### PROBLEMA DUNLE DI Wolfe

Sono  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  e  $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$  convesse e differenziabili e sia  $x^* \in \mathbb{R}^n$  una soluzione ottimale del problema:

$$\begin{aligned} & \min_x f(x) \\ & \text{s.t. } g(x) \leq 0 \end{aligned}$$

S.I.  $\mu^* \in \mathbb{R}^m$  è un VETTORE DI MULIPPLICATORI TAU CHE LA COPPIA  $(x^*, \mu^*)$  SODDISFA LE  
KKT. ALLORA  $(x^*, \mu^*)$  È UNA SOLUZIONE OTTIMALE DEL DUALE DI Wolfe:

$\curvearrowright$  LAGRANGEAN

$$\boxed{\max_{x, \mu} \mathcal{L}(x, \mu) = f(x) + \mu^T g(x)}$$

- $\mu \geq 0$
- $\nabla_x \mathcal{L}(x, \mu) = 0$

Dim:

S.I. HA CHE LE KKT CONSIDERANO:

$$\cdot \mu^* \geq 0$$

$$\cdot g(x^*) \leq 0$$

$$\cdot \mu_i^* g_i(x^*) = 0$$

$$\cdot \nabla_x \mathcal{L}(x^*, \mu^*) = 0$$

$$\cdot \mathcal{L}(x^*, \mu^*) = f(x^*) + \sum_i \mu_i^* g_i(x^*) = \boxed{f(x^*)}$$

Dopo che  $\mu_i^* g_i(x^*) = 0$

S.I.  $(x, \mu)$  È UNA GENERICA SOLUZIONE AMMISIBILE PER IL DUALE DI WOLFE.

POICHE'  $g_i(x^*) \leq 0 \forall i$  E  $\mu_i \geq 0 \forall i$  OTENGO:

$$\mathcal{L}(x^*, \mu^*) = f(x^*) \geq f(x^*) + \sum_i \mu_i^* g_i(x^*) = \boxed{\mathcal{L}(x^*, \mu)}$$

ma  $\mathcal{L}(x, \mu)$  è convessa rispetto ad  $x$  poiché è somma di funzioni convesse (lineari, con coefficienti positivi), delle funzioni convesse  $f, g_1, \dots, g_m$

Meno:

$$\mathcal{L}(x^*, \mu) \geq \mathcal{L}(x, \mu) + \nabla_x \mathcal{L}(x, \mu)(x^* - x)$$

ma  $\nabla_x \mathcal{L}(x, \mu) = 0$  poiché è soluzione ottimale ammessa per il duali, si ha

$$\mathcal{L}(x^*, \mu^*) \geq \mathcal{L}(x^*, \mu) \geq \mathcal{L}(x, \mu)$$

ma essendo che  $\mathcal{L}(x, \mu)$  è una funzione soluzioni ammesse per il duali, se  $(x^*, \mu^*)$  è soluzione ammessa per il duali, è quindi  $(x^*, \mu^*)$  è ottimale per il duali! □

TRAMANDO AL PROBLEMA DOPPIATO DELLO SVM, SI HA CHE QUESTO È CONVESO CON VINCULANTI, E INOLTRE LE HMT RISULTANO ESSERE C.N.S DI OTTIMIZZAZIONE.

↳ poiché il problema è convesso!

puoi se  $(w^*, b^*, \epsilon^*)$  è soluzione ottimale, allora esistono  $\lambda^* \in \mathbb{R}$  e  $\mu^* \in \mathbb{R}^m$  tali che:

$$(w^*, b^*, \epsilon^*, \lambda^*, \mu^*) \text{ soddisfano HMT}$$

puoi posso scrivere il duali di wpe come:

$$g_1(x) + \dots + g_m(x)$$

$$\max \mathcal{L}(w, b, \xi, \alpha, \mu)$$

•  $\frac{\partial \mathcal{L}}{\partial w} = 0$

•  $\frac{\partial \mathcal{L}}{\partial b} = 0$

$$\nabla_w \mathcal{L}(w, b, \xi, \alpha, \mu) = 0 \quad (1)$$

$$\nabla_b \mathcal{L}(w, b, \xi, \alpha, \mu) = 0 \quad (2)$$

$$\nabla_\xi \mathcal{L}(w, b, \xi, \alpha, \mu) = 0 \quad (3)$$

Dove

$$\mathcal{L} = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i + \sum_i \mu_i (-\xi_i) + \sum_i \alpha_i (1 - y^{(i)}(w^\top x^{(i)} + b)) - \xi_i$$

Dove il vincolo su  $\xi$  è stato posto in forma standard ( $-\xi_i \leq 0$ )

possiamo aggiungere i vincoli:

$$(1) \quad \nabla_w \mathcal{L}(\cdot) = 0 \Rightarrow w - \sum_{i=1}^n \alpha_i y^{(i)} x^{(i)} = 0 \Rightarrow w = \sum_{i=1}^n \alpha_i y^{(i)} x^{(i)}$$

$$(2) \quad \nabla_b \mathcal{L}(\cdot) = 0 \Rightarrow - \sum_{i=1}^n \alpha_i y^{(i)} = 0 \Rightarrow \sum_{i=1}^n \alpha_i y^{(i)} = 0$$

$$(3) \quad \frac{\partial \mathcal{L}(\cdot)}{\partial \xi_i} = 0 \Rightarrow C - \mu_i + \alpha_i = 0 \quad \forall i \Rightarrow \alpha_i = C - \mu_i$$

Possiamo allora sostituire i valori ricavati ottenendo:

$$\mathcal{L}(w, \alpha, \mu) = \frac{1}{2} \|w\|^2 + \sum_i \alpha_i - \sum_i \alpha_i y^{(i)} w^\top x^{(i)}$$

E POSSIAMO SCRIVERE IL LOSS FUNCTION DI  $\alpha$ :

$$\begin{aligned} \mathcal{L}(w, \alpha) &= \frac{1}{2} w^T w + \sum_i \alpha_i - \sum_i \alpha_i y^{(i)} w^T x^{(i)} \\ &= \frac{1}{2} w^T w + \sum_i \alpha_i - w^T \sum_i \alpha_i y^{(i)} x^{(i)} \\ &= \sum_i \alpha_i + w^T \left( \frac{1}{2} w - \sum_i \alpha_i y^{(i)} x^{(i)} \right) \end{aligned}$$

$$\Rightarrow w = \sum_i \alpha_i y^{(i)}$$

Allora:  $\mathcal{L}(w, \alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)} x^{(j)} = \alpha^T Q \alpha - \frac{1}{2} Q^T Q \alpha$

con  $Q_{ij} = y^{(i)} y^{(j)} x^{(i)} x^{(j)}$ .

IL PROBLEMA DIVENTA ALLORA:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \alpha^T Q \alpha + C \alpha \\ \text{s.t.} \quad & 0 \leq \alpha \leq C \\ & \alpha^T y = 0 \end{aligned}$$

ED ESSENDO CHE IL PROBLEMA È QUADRATICO, STrettAMENTE CONVesso E CON VINCOLI LINEARI, POSSO MINIMIZZARE TUTTI GLI ALTRI PARAMETRI MISCIANDO UNICO VINCONE  $\alpha^*$ :

$$w^* = \sum_{i=1}^N \alpha_i^* y^{(i)} x^{(i)}$$

$$\mu_i^* = C - \alpha_i^*$$

$$b^* = \frac{y^{(i)}}{(w^*)^T x^{(i)}} \quad \forall i \text{ t.c. } \alpha_i \in (0, C)$$

$$\xi_i^* = \max \left\{ 0, 1 - y^{(i)} ((w^*)^T x^{(i)} + b^*) \right\}$$

OSSERViamo:

$$\alpha_i^* (1 - y^{(i)} ((w^*)^T x^{(i)} + b^*) - \xi_i^*) = 0$$

$$\mu_i^* \xi_i^* = (C - \alpha_i^*) \xi_i^* = 0$$

Allora SG:

- $\alpha_i^* \in (0, C)$ : Allora  $\xi_i^* = 0$  e quindi  $y^{(i)} ((w^*)^T x^{(i)} + b^*) = 1$ , ovvero 1 PUNTO  
Si trova ESATTAMENTE SULLA SUPERFICIE DI SEPARAZIONE. E' PUNTO  
 $x^{(i)}$  SARA' UN VETTORE DI SUPPORTO!

- $\alpha_i^* = 0$ :  $\xi_i^* = 0$  ma  $y^{(i)} ((w^*)^T x^{(i)} + b^*) \geq 1$ , ovvero  $x^{(i)}$  non e' dentro ma sia  
un VETTORE DI SUPPORTO

- $\alpha_i^* = C$ : Poco ESSERE  $\xi_i^* > 0$ , allora  $1 - y^{(i)} ((w^*)^T x^{(i)} + b^*) = \xi_i^* > 0$

QUINDI IL PROBLEMA DUELE E' CONDOTTO DA OTTIMIZZARE POICHE' I VINCOLI SONO PIU'  
SEMPLICI DA OBTENERE. INVECE LA MATRICE Q E' una MATRICE DI SIMMETRIA tra i campioni.

TUTTAVIA POSSIAMO UTILIZZARE METODI DI SIMPLIFICAZIONE ALTRERUNI (E PIU' POTENTI) RISPETTO AL SEMPLICE

## Prodotto scalare

SI USANO QUINDI LE FUNZIONI KERNEL  $K: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  CHE O PERMETTONO DI INCAPSULARE IL PRODOTTO SCALARE:

$$Q_{ij} = y_i y_j K(x_i, x_j)$$

$$h(x) = \sum_{i=1}^m a_i y^{(i)} K(x^{(i)}, x) + b$$

PER TUTTI GLI ELEMENTI DEL DATASET DEVIOSI DEVONO ESSERE VOLTI, OVVERO SE  $\mathbf{X}$  DÀ UN SET DI MATERICI

$Q_{ij} = y_i y_j K(x_i, x_j)$  È SEMPER DEFINITA POSITIVA, OVVERO SE E SOLO SE  $K$  MANTIENE IL

PRODOTTO SCALARE DELLO SPAZIO INIZIALE IN UNO SPAZIO A DIMENSIONE MAGGIORA.

QUINDI  $K$  È VOLTO, DATI  $u, v \in \mathbb{R}^n$ , SE  $\exists \phi: \mathbb{R}^n \rightarrow \mathbb{R}^m$   $m > n$  TAU ANTI:

$$u, v \in \mathbb{R}^n \xrightarrow{\phi} \phi(u), \phi(v) \in \mathbb{R}^m \Rightarrow \phi(u)^T \phi(v) = K(u, v)$$

$\Rightarrow$  IL KERNEL PERMETTE DI COSTRUIRE CLASSIFICATORI NON LINEARI POICHÉ CIO' CHE G' È  
LINEARE IN  $K(\mathbb{R}^n, \mathbb{R}^m)$  PUÒ NON ESSERE IN  $\mathbb{R}^n \times \mathbb{R}^m$

METHODI DI DECOMPOSIZIONE

IN CONGRICO PROBLEMI

$$\min_{x \in S \subseteq \mathbb{R}^n} f(x)$$

$$\min_{x_1, x_2, \dots, x_n} f(x_1, \dots, x_n)$$

$$x_m \in X_M \subset \mathbb{R}^{m_m}$$

IN QUESTO POSSIAMO SCOMPONERE LE VARIABILI DEL PROBLEMA IN M GRUPPI OTTENENDO QUINDI DEI SOTTOPROBLEMI DELLA FORMA

$$f(\bar{x}_1, \dots, \bar{x}_i, \dots, \bar{x}_M)$$

$$x_i \in X^M_i$$

QUESTO PROCESSO DI SCOMPONENTIZZAZIONE RISULTA OSSERVANDO UNICO PRIMO IL NUMERO DI VARIABILI È GRANDE, PRIMO  $f$  RISULTA SEPARABILE E CONVESSA RISPETTO AI SINGOLI SOTTOBLOCCHI.

### METODI SEQUENZIALI (GAUSS-SEIDEL)

L'AGGIORNAMENTO È DEL TIPO

$$x_i^{k+1} = \underset{x_i \in S_i}{\operatorname{arg\,min}} f(x_1^k, \dots, \bar{x}_{i-1}^k, x_i, \bar{x}_{i+1}^k, \dots, \bar{x}_M^k)$$

IL METODO CONVERGE A PUNTI STABILMENTE SE  $f$  È CONVESSA, OPPURE SE  $f$  È SISTEMATICAMENTE CONVESSA PER LE COMPONENTI SCELGENDO  $x_i$  CON  $i=1, \dots, M$ . IN QUESTO BISOGNA È SE M>2.

### METODI PARALLELI (JACOBI)

$$x_i^{k+1} = \underset{x_i \in S_i}{\operatorname{arg\,min}} f(x_1^k, \dots, \bar{x}_{i-1}^k, x_i, \bar{x}_{i+1}^k, \dots, \bar{x}_M^k)$$

DRO DRO:  $x^{k+1} = \underset{x \in S}{\operatorname{arg\,min}} f(x)$

$$x = (x_1, \dots, c_i, \dots, x_n)$$

$$i=1, \dots, n$$

## METHOD CON SOTTOPPOSIZIONE A BLOCCHI

AD OGNI ITERAZIONE SI SCEGLIE UN WORKING SET  $W_h \subseteq \{1, \dots, n\}$ .

L'AGLORITMICO E' PURO OTTENUTO DA:

$$\tilde{x}_{w_h}^{h+1} = \underset{x_{w_h}}{\operatorname{argmin}} \{ (x_{w_h}, x_{\bar{w}_h}^h) \}$$

► VERSIONE SCOTT DEL  
WORKING SET

Dove  $\bar{W}_h = \{1, \dots, n\} \setminus W_h$

Ness

$$x_i^{h+1} = \begin{cases} x_i^h & \text{se } i \in \bar{W}_h \\ \tilde{x}_i^{h+1} & \text{se } i \in W_h \end{cases}$$

LA CONVERGENZA DELLO SCHEMA DIPENDE DALLA REGOLE DI SELEZIONE DI  $W_h$ . SI HANNO DUE REGOLE:

- Regole classiche:

$$\exists M > 0 \text{ tc: } \forall \{1, \dots, n\} \ni t_{h=0, 1, \dots} \quad \exists l(h) \leq M \text{ tc } i \in W_{h+l(h)}$$

- Regole di Gauss-Southwell:

$$\forall h \exists i(h) \in W_h \text{ tc: } \left| \frac{\partial f(x^h)}{\partial x_{i(h)}} \right| \geq \left| \frac{\partial f(x^h)}{\partial x_j} \right| \quad \forall j \in \{1, \dots, n\}$$

ONDE SI INSERISCE NEL WORKING SET LA VARIABILE A MASSIMO DISCESSO.

### ALGORITMO DI DECOSTRUZIONE PGS SUM

DUAL SUM È DATO DA

$$\begin{aligned} \min_{\alpha} & \frac{1}{2} \alpha^T Q \alpha - c^T \alpha \\ \text{s.t.} & 0 \leq \alpha \leq C \\ & \alpha^T y = 0 \end{aligned}$$

LA COMPLICATI<sup>Y</sup> DI QUESTO PROBLEMA RISIDE NEL Fatto CHE IL VINCERE UNA DI  
OCCHIANDA NON È SEPARABILE ED INOLTRE Q È UNA MATRICE DENS<sup>A</sup> CHE SCALA  
MOLTO AL CRESCERE DEL NUMERO DI ESEMPI.

VOCEREMO PUNTO DECOSTRUZIONE IL PROBLEMA INTRODUCENDO UN WORKING SET

$\bar{W} = \{1, \dots, m\}$ . IL SOTTOPROBLEMA ALL'ITERAZIONE  $k$  SARÀ PUNTO DATO DA:

$$\min_{\alpha_W} \frac{1}{2} \alpha_W^T Q_{WW} \alpha_W + \alpha_W^T (Q_{W\bar{W}} \alpha_{\bar{W}} - c)$$

$$0 \leq \alpha_W \leq C$$

$$\alpha_W^T y_W = -(\alpha_{\bar{W}}^T)^T y_{\bar{W}}$$



$$\alpha^T y = 0 \Rightarrow \alpha_W^T y_W + \alpha_{\bar{W}}^T y_{\bar{W}} = 0$$

PERché VARIABILI DOVRAO SCUVERE NEL WORKING SET? SE SI, SCUVERE UNA SOLO  
VARIABILE  $w \in \{i\}$  SI HA CHE:

$$\alpha_i y_i = - \sum_{j \neq i} \alpha_j^T y_j = b \rightarrow \alpha_i = \frac{b}{y_i}$$

E' POSSIBILE SCARICARE DUE SEMPRE LO STESSO DISCUSSO. SI SCEGLIONO quindi DUE VARIABILI  $w = \{i, j\}$  IN MODO CHE IL SOTOPROBLEMA SIA RESOLVIBILE IN FORMA CAVIUS.

PUNTO CON  $w = \{i, j\}$  OTTENUTO

$$(x_i, x_j) \begin{pmatrix} \varphi_{ii} & \varphi_{ij} \\ \varphi_{ji} & \varphi_{jj} \end{pmatrix} - x_i - x_j + p^T \begin{pmatrix} x_i \\ x_j \end{pmatrix}$$

(Punto - e)

SOGGETTO AI VINCOLI

$$0 \leq x_i, x_j \leq C$$

$$x_i y_i + x_j y_j = - \sum_{h \neq i, j} x_h y_h$$

I METODI COSÌ COSTRUITI SONO DETTI SEQUENTIAL MINIMAL OPTIMIZATION. POSSONO DEFINIRE IN CUI MODO SCENDERE QUESTE DUE VARIABILI DEL WORKING SET.

DURANTE UNA DISSONZA VOGLIAMO OBTENERE UNA DIREZIONE DI DISCESA AMMISIBILE E CON SOLO DUE COORDINATI NON NULLI:  $d^{ij} = (0 \ 0 \ \dots \ d_i \ 0 \ \dots \ 0)$

PROPOSIZIONE:

SIA  $\bar{x} \in S = \{0 \leq x \leq C, x^T y = 0\}$ . L'INSIEME DELLE DIREZIONI AMMISIBILI IN  $\bar{x}$  È DATO DA

DA  $D(\bar{x}) = \{d \in \mathbb{R}^n \mid d^T y = 0, d_i > 0 \forall i \in L(\bar{x}) \text{ E } d_i \leq 0 \forall i \in U(\bar{x})\}$

con  $L(\bar{x}) = \{i \mid \bar{x}_i = 0\}$  E  $U(\bar{x}) = \{i \mid \bar{x}_i = C\}$

VOGLIAMO OBTENERE UNA DIREZIONE  $d^{ij}$  AMMISIBILE. SIA  $(d^{ij})^T y = 0$

$\Rightarrow d_i y_i + d_j y_j = 0$  ovvero

$$d_i = \frac{1}{y_i}$$

$$d_j = \frac{1}{y_j}$$

SCENARIO I

SCENARIO J

.  $i \in L(\alpha^u)$  solo se  $y_i = 1$

.  $i \in U(\alpha^u)$  solo se  $y_i = -1$

.  $0 \leq \alpha_i^u \leq c$  oh  $\cancel{y_i}$

.  $j \in L(\alpha^u)$  solo se  $y_j = 1$

.  $j \in U(\alpha^u)$  solo se  $y_j = -1$

.  $0 \leq \alpha_j^u \leq c$  oh  $\cancel{y_j}$

CONSIDERAMOS ORA:

$$L(\alpha) = L^+(\alpha) \cup L^-(\alpha) = \left\{ i \in L(\alpha) \mid y_i = 1 \right\} \cup \left\{ i \in L(\alpha) \mid y_i = -1 \right\}$$

$$U(\alpha) = U^+(\alpha) \cup U^-(\alpha) = \left\{ i \in U(\alpha) \mid y_i = 1 \right\} \cup \left\{ i \in U(\alpha) \mid y_i = -1 \right\}$$

PROPOSIZIONE:

UN DIRIGIBILITÀ  $d^{10} = \{ 0 \quad 0 \quad 0 \quad \frac{1}{x_1} \quad 0 \quad 0 \quad \frac{1}{x_2} \quad 0 \quad 0 \}$  È AMMISIBILE IN UN PUNTO  $\alpha^u$

SE E SOLO SE  $i \in R(\alpha^u) = L^+(\alpha^u) \cup U^-(\alpha^u) \cup \{ i \mid 0 \leq \alpha_i^u \leq c \}$  E

$\exists \alpha \in S(\alpha^u) = L^-(\alpha^u) \cup U^+(\alpha^u) \cup \{ i \mid 0 \leq \alpha_i^u \leq c \}$

D.m.

SUPPONIAMO  $d^{\text{ij}}$  AMMISSIBILE E SUPPONIAMO, PER ASSERIRE, CHE  $\bar{x} \in U(\bar{x}^n)$ .  
 ALLORA SE HA CHE  $d_{\bar{x}} = -\frac{1}{x_j} = -\frac{1}{-1} = 1 > 0$  MA  $\bar{x}_j^n \in C$  QUINDI DVE ESSERE CHE  
 $d_{\bar{x}} \leq 0 \rightarrow$  ASSERIRE

VICENDA, SUPPONIAMO CHE  $\bar{x} \in L^+(\bar{x}^n) \subset \bar{x} \in U^+(\bar{x}^n)$ , ALLORA

$$y^T d^{\text{ij}} = y_i \cdot \frac{1}{x_i} + y_j \left( -\frac{1}{x_j} \right) = 1 - 1 = 0$$

DVE ESSERE  $d_{\bar{x}} \geq 0$ , ALLORA  $d_{\bar{x}} = -\frac{1}{x_j} = \frac{+1}{+1} = +1 > 0 \checkmark$

DVE ESSERE  $d_{\bar{x}} \leq 0$ , ALLORA  $d_{\bar{x}} = -\frac{1}{x_j} = \frac{-1}{+1} = -1 < 0 \checkmark$

QUINDI  $d^{\text{ij}}$  È AMMISSIBILE SE  $\bar{x} \in R(\bar{x}^n) \subset \bar{x} \in S(\bar{x}^n)$  □

PROPOSIZIONE:

$d^{\text{ij}}$  È DI DISCESA IN  $\bar{x}^n \Leftrightarrow$

$$\boxed{\frac{\nabla_i f(\bar{x}^n)}{x_i} \leq \frac{\nabla_j f(\bar{x}^n)}{x_j}}$$

D.m.:

$f$  È CONVEXA  $\Rightarrow d^{\text{ij}}$  È DI DISCESA SE E SOLO SE  $\nabla f(\bar{x}^n)^T d^{\text{ij}} \leq 0$

$$\text{MA } \nabla f(\bar{x}^n)^T d^{\text{ij}} = \sum_h \underbrace{\frac{\partial f(\bar{x}^n)}{\partial x_h} d_h}_{\text{Dove } h \neq i, j} = \underbrace{\frac{\partial f(\bar{x}^n)}{\partial x_i} d_i}_{\text{O } h=i} + \underbrace{\frac{\partial f(\bar{x}^n)}{\partial x_j} d_j}_{d_j \leq 0}$$

$$\text{ma } d_i = \frac{1}{x_i} \in d_j = \frac{1}{x_j} \text{ stessa } \frac{1}{x_i} \frac{\partial f(x^u)}{\partial x_i} - \frac{1}{x_j} \frac{\partial f(x^u)}{\partial x_j} < 0 \quad \square$$

NBBIAMO PENSATO OTTENUTO CHE DOBBIAMO SCHEGGERE

$$W = \{i, j\} \text{ con}$$

$$i \in R(x^u)$$

$$j \in S(x^u)$$

d'AMMISIBILE

TOUCH

$$\frac{\nabla_i f(x^u)}{x_i} < \frac{\nabla_j f(x^u)}{x_j}$$

id di  
discussione

Algoritmo Smo:

$$\text{DATI } P, x^0 \in \mathbb{R}^n, k=0 \text{ E } \boxed{\nabla f(x^0) = P x^0 - c = -c \text{ SE } \text{PENSANDO } x^0 = 0}$$

WHILE (Criterio di arresto non soddisfatto).

. SCHEGLIO  $W = \{i, j\}$  CON  $i \in R(x^u), j \in S(x^u)$  TOUCH

$$\frac{\nabla_i f(x^u)}{x_i} < \frac{\nabla_j f(x^u)}{x_j}$$

. CALCOLO  $x_i^* \in x_j^*$  USANDO IL SOTTEPROBLEMA IN DUE VARIABILI

. ACCORDO

$$x_h^u = \begin{cases} x_i^* & h=i \\ x_j^* & h=j \\ x_h^u & \text{altrimenti} \end{cases}$$

. CALCOLA  $\nabla f(x^{u+1}) = \nabla f(x^u) + P_i (x_i^{u+1} - x_i^u) + P_j (x_j^{u+1} - x_j^u)$

$$\cdot \quad H = H + 1$$

END WHILE

Si nota che l'incremento del gradiente risulta di ceduta soltanto

$$Q_i \text{ e } Q_j$$

### PROPSITI

$\bar{x}^*$  È un punto di minimo globale per il doppio di SVM SG E solo SG:

$$\boxed{\max_{i \in R(\bar{x}^*)} \left\{ \frac{-\nabla_i f(\bar{x}^*)}{x_i} \right\} \leq \min_{j \in S(\bar{x}^*)} \left\{ \frac{-\nabla_j f(\bar{x}^*)}{x_j} \right\}}$$

Dim:

Il problema è convesso con vincoli lineari, quindi le MHT sono condizioni necessarie e sufficienti di ottimalità.

Passiamo quindi i vincoli del doppio in forma standard:

$$\begin{aligned} & \underline{x}_i = 0 \rightsquigarrow \lambda_i \mathbb{M} \\ \alpha_i \in [0, 1] \quad & -\underline{x}_i \leq 0 \rightsquigarrow \mu_i \mathbb{M}^m \\ & \underline{x}_i - C \leq 0 \rightsquigarrow \mu_i^+ \mathbb{M}^m \end{aligned}$$

scriviamo ora le MHT:

$$\mu_i^+ \geq 0, \mu_i^- \geq 0 \quad \forall i$$

$$\cdot \mu_i^+ (\alpha_i - c) = 0 \quad \forall i$$

$$\cdot \mu_i^- (\alpha_i) = 0 \quad \forall i$$

$$\cdot \nabla_{\alpha} \mathcal{L}(\alpha, \lambda, \mu^+, \mu^-) = 0$$

Dove la costante vale:  $\mathcal{L}(\alpha, \lambda, \mu^+, \mu^-) = f(\alpha) + \lambda(y^T \alpha) + \sum_{i=1}^n \mu_i^- (-\alpha_i) + \sum_{i=1}^n \mu_i^+ (\alpha_i - c)$

CALCOLIAMO ADesso IL GRADIENTE:

$$\Rightarrow \nabla_{\alpha} \mathcal{L}(") = \nabla_{\alpha} [f(\alpha) + \lambda y_i - \mu_i^+ + \mu_i^-] = 0 \quad \forall i$$

dove  $\boxed{\nabla_{\alpha} [f(\alpha) + \lambda y_i] = -\mu_i^+ + \mu_i^-}$

$$\Rightarrow \nabla_{\alpha} [f(\alpha) + \lambda y_i] \left\{ \begin{array}{l} = 0 \quad \text{SE } \underline{\alpha_i < C} \\ = -\mu_i^+ \leq 0 \quad \text{SE } \underline{\alpha_i = C} \\ = -\mu_i^- \geq 0 \quad \text{SE } \underline{\alpha_i = 0} \end{array} \right.$$

DIVISO NUOVA PER  $y_i$ :

$$\frac{\nabla_{\alpha} [f(\alpha) + \lambda y_i]}{y_i} \left\{ \begin{array}{l} = 0 \quad \text{SE } \underline{\alpha_i < C} \\ \leq 0 \quad \text{SE } i \in \underline{[U^+(\alpha) \cup L^-(\alpha)]} \\ \geq 0 \quad \text{SE } i \in \underline{[U^-(\alpha) \cup L^+(\alpha)]} \end{array} \right.$$

OVVGEZO

$$\frac{\nabla_i f(\alpha)}{y_i} \begin{cases} = -\lambda & \text{SE } 0 < \alpha_i < C \\ \leq -\lambda & \text{SE } i \in [U^+(\alpha) \cup L^-(\alpha)] \\ \geq -\lambda & \text{SE } i \in [U^-(\alpha) \cup L^+(\alpha)] \end{cases}$$

punto) SE  $\bigcup_{j \in [U^+(\alpha) \cup L^-(\alpha)]} \{j \mid 0 < \alpha_j < C\} = S(\alpha)$  si no:

$$\boxed{\frac{\nabla_j f(\alpha)}{y_j} \leq -\lambda}$$

punto) SE  $\bigcup_{i \in [U^-(\alpha) \cup L^+(\alpha)]} \{i \mid 0 < \alpha_i < C\} = R(\alpha)$  si no:

$$\boxed{\frac{\nabla_i f(\alpha)}{y_i} \geq -\lambda}$$

punto:

$$\frac{\nabla_i f(\alpha)}{y_i} \geq -\lambda \geq \frac{\nabla_j f(\alpha)}{y_j} \quad \forall i \in R(\alpha) \in \boxed{\nabla_j f(\alpha) \in S(\alpha)}$$

caso medio / segn:

$$\boxed{-\frac{\nabla_i f(\alpha)}{y_i} \leq \lambda \leq -\frac{\nabla_j f(\alpha)}{y_j} \quad \forall i \in R(\alpha) \in \nabla_j f(\alpha) \in S(\alpha)}$$

$y_i$

$y_j$

$$\Rightarrow -\frac{\nabla_i f(\alpha)}{y_i} \leq -\frac{\nabla_j f(\alpha)}{y_j} \quad \forall i \in R(\alpha) \in \forall j \in S(\alpha)$$

E quindi, se maggior ragione, vale così

$$\max_{h \in R(\alpha)} \left\{ -\frac{\nabla_h f(\alpha)}{y_h} \right\} \leq \min_{h \in S(\alpha)} \left\{ -\frac{\nabla_h f(\alpha)}{y_h} \right\}$$

□

cordano:

SIA  $\alpha$  non ottimale, allora si avrà che

$$\max_{h \in R(\alpha)} \left\{ -\frac{\nabla_h f(\alpha)}{y_h} \right\} > \min_{h \in S(\alpha)} \left\{ -\frac{\nabla_h f(\alpha)}{y_h} \right\}$$

DEFINIZIONE:

una coppia  $(i^*, j^*)$  è detta most violating pair se:

$$i^* \in \operatorname{argmax} \left\{ -\frac{\nabla_i f(\alpha)}{y_i} \right\}$$

$$j^* \in \operatorname{argmin} \left\{ -\frac{\nabla_j f(\alpha)}{y_j} \right\}$$

$$\left\{ \begin{array}{l} \alpha_{hR(\alpha)} \\ \alpha_{hS(\alpha)} \end{array} \right\} \quad \left\{ \begin{array}{l} y_h \\ y_b \end{array} \right\}$$

$$\left\{ \begin{array}{l} \alpha_{hS(\alpha)} \\ \alpha_{hR(\alpha)} \end{array} \right\} \quad \left\{ \begin{array}{l} y_b \\ y_h \end{array} \right\}$$

perciò  $\bar{f}_{iR(\alpha)} \in \bar{f}_{jS(\alpha)}$  TAU CHE:

$$\frac{-\nabla_i f(\alpha)}{y_i} > \frac{-\nabla_j f(\alpha)}{y_j}$$

MENO:

$$\frac{\nabla_i f(\alpha)}{y_i} < \frac{\nabla_j f(\alpha)}{y_j}$$

→ È PROPOSSIBILE CONDIZIONI PRESENTE  
NELL'ALGORITMO SNO!

PROPOSIZIONE:

SIA  $\left\{ \alpha^k \right\}$  PRODOTTA DALL'ALGORITMO DI SNO SEGUENDO, AD OGNI  
ITERAZIONE, UNA MOST VIOLENTING PNR CON IL WORKING SET, ALLORA:

$$\lim_{k \rightarrow \infty} \alpha^k = \alpha^*$$

con  $\alpha^*$  PUNTO DI MINIMO CROSPOLIG

## DEFINITION 02:

$$\max_{h \in R(\alpha)} \left\{ -\frac{\nabla_h f(\alpha)}{\gamma_h} \right\}$$

$$\min_{h \in S(\alpha)} \left\{ -\frac{\nabla h f(\alpha)}{\gamma_h} \right\}$$

$M(\omega)$

$$m(\alpha)$$

PER  $\alpha^*$  OTTIMO SAPPiamo CHE  $M(\alpha^*) \leq m(\alpha^*)$

PER PRIMO NUOVO IL COTERZO DI ARZEGGIO DI SMO?

Si considera  $M(x^n) \leq m(x^n) + \varepsilon$ . Tuttavia, essendo una pista

RELACIONES ENTRE SEMPRE PRESENTE CONTINUO (POUCHE R, S, UNA SOLO  
INSTANCIA CONTINUA) ALGOZAR NO SE DEDICA CUE SI QUITA AL CRITERIO

DI ANASTO; PERO, SE PUESTA A VERIFICAR, ALGO VERO, EN UN NUEVO

FINITO DI PASSI, SI AVRA' CHE  $M(2^4) - m(2^4) \leq \varepsilon$

## SVM (WB/ASED)

Si va a ~~no~~ considerar el BIS.

$$\min_w \frac{1}{2} \|w\|^2 + \sum_{i=1}^N \max \left\{ 0, 1 - y^{(i)} (w^T x^{(i)}) \right\}$$

DOLG AL PESO DEL BIAS  $b$ ; SI VA A PARTE AUS FEATURE COSTANTE PER TO ANS PER OGNI CAMPIONE! IN QUESTO MODO SI HA UNA MIGLIOR RECUPERAZIONE.

SE NOI CONSIDERIAMO IL PROBLEMA DUALE:

$$\min_z \frac{1}{2} z^T \Phi z - c^T z$$

$$0 \leq z_i \leq C$$

→ SPARISCE IL VINCULO UNEDE  
DI KRUEGERA!

SI RISOLVE TRAMITE IL DUAL COORDINATE DESCENT IL PESO AGGIORNAMENTO E' SOLO VARIABILE NELLA ZETA. LA SECUZA DEGLI VARIABILI PUO'ESSERE SEQUENZIALE OPPURE CASUALE. SE SIANO NEL CASO DI KERREL UNIVERSE ALLORA NON OCCORRE CONSIDERARE LA MATRICE  $\Phi$

SE SI HANNO DATI SET GRADO ALTO SI GLI SI PENSANO CO RISPEZZATO IN modo DA CONSIDERARE LA DIFFERENZIABILITA'

$$\min_w \frac{1}{2} \|w\|^2 + \sum_{i=1}^N \max \left\{ 0, 1 - y^{(i)} (w^T x^{(i)}) \right\}^2$$

## PROBLEMI DI SCARICO FINITO

Sono problemi DGL TIPO

$$\min_{x \in \mathbb{R}^m} f(x) = \frac{1}{N} \sum_{i=1}^N f_i(x)$$

Dove  $f$  è privo di media dei TERMINI. Si risolve utilizzando il METODO DI TIPO CONIGLIO STOCHASTICO:

$$x^{H+1} = x^H + \Delta_H d_H$$

$$d_H = \frac{1}{|\beta|} \sum_{i \in \beta} -\nabla f_i(x^H)$$

1 MINIBATCH

$$\beta \subseteq \{1, \dots, N\}$$

Mentre il passo  $\Delta_H$  è tipicamente scelto da una sequenza decrescente prefissata.

notiamo che:

se  $|\beta|=1$  e  $\beta_H = \{i_H\}$  con  $H$  scelta in modo uniforme, ovvero

$$p_i = \frac{1}{N} \forall i, \text{ allora:}$$

$$\begin{aligned} E[d_H] &= E[-\nabla f_{i_H}(x^H)] \\ &\stackrel{\text{DEFINIZIONE DI}}{=} \sum_{i=1}^N p_i (-\nabla f_i(x^H)) \\ &= \frac{1}{N} \sum_i -\nabla f_i(x^H) \end{aligned}$$

dove  $-\nabla f(x^H) \Rightarrow$  APPROXIMA L'ANTIGRADIENTE GUARDANDO SOLO UNA SELEZIONE COMPLEMENTARE, in modo, nel nostro caso il valore effettivo dell'antigradiente!

SGD

Proposizioni:

SUPponiamo che sia  $\|\nabla f_i(x)\| \leq L$  per tutti  $i, x \in \mathbb{R}^m$ . Sia inoltre  $\nabla f_i$  LIPSCHITZ-CONTINUA. ASSUMO CHE LA SEQUENZA  $\{\alpha_n\}$  sia tale CHE:

$$\sum_{t=0}^{\infty} \alpha_t = +\infty \quad ; \quad \sum_{t=0}^{\infty} \alpha_t^2 < +\infty$$

E CHE AD Ogni ITERAZIONE L'ALGORITMO DÀ IN OUTPUT LA SOLUZIONE  $\bar{x}^{k+1}$  DONGE  $\bar{x}^{k+1} = x^r$  PER qualcHE  $r = 0, -1, k$  DONGE:

$$P(r, t) = \frac{\alpha_t}{\sum_{i=0}^t \alpha_i}$$

allora

$$\lim_{k \rightarrow \infty} \mathbb{E} \left[ \|\nabla f(\bar{x}^k)\|^2 \right] = 0$$

$\hookrightarrow$  IN PRACTICE, ALLA CONVERGENZA, LE SOLUZIONI TENDONO A ASSOMIGLIARSI.

$$f(x^{k+1}) \leq f(x^k) + \frac{\alpha_k^2 L^2}{2} - \alpha_k \nabla f(x^k)^T \nabla f(x^k)$$

$\Rightarrow$  non abbiamo la certezza di decrescere i costi iterativi e  
l'iterazione può si passare a indefinito con il valore atteso:

$$\sum_{t=0}^n \alpha_t \mathbb{E} [\|\nabla f(x^t)\|^2] \leq M + \frac{LG^2}{2} \sum_{t=0}^n \alpha_t^2$$

$$\rightarrow \mathbb{E} [\|\nabla f(z^{t+1})\|^2] = \sum_{t=0}^n \mathbb{E} [\|\nabla f(x^t)\|^2] \cdot \underbrace{\mathbb{P}(z^{t+1} = x^t)}_{\frac{\alpha_t}{\sum_{i=0}^n \alpha_i}} = *$$

$$* = \frac{1}{\sum_{i=0}^n \alpha_i} \cdot \sum_{t=0}^n \alpha_t \cdot \mathbb{E} [\|\nabla f(x^t)\|^2] \leq M + \frac{LG^2}{2} \sum_{t=0}^n \alpha_t^2 = \infty$$

$$\rightarrow \mathbb{E} [\|\nabla f(z^{t+1})\|^2] \leq M + \underbrace{\frac{LG^2}{2} \sum_{t=0}^n \alpha_t^2}_{\hookrightarrow 0} \hookrightarrow \infty$$

punto, ad esempio, la sequenza  $\alpha_n = \frac{\alpha_0}{n}$  smentisce l'asserzione!

Analizziamo adesso la complessità:

METHODO

Caso Convesso

Caso Funzione Convessa

	$O(\sqrt{\epsilon})$	$O(N \log \frac{1}{\epsilon})$
GD	$O(\sqrt{\epsilon^2})$	$O(\sqrt{\epsilon})$

## MINIBATCH SGD

DATI  $X \in \mathbb{R}^n$ ,  $f(x) = \frac{1}{N} \sum_{i=1}^N f_i(x)$ ,  $H=0$

WHILE ( criterio di arresto non soddisfatto):

- DIVIDIAMO  $\{1, \dots, N\}$  in M Blocco Disgiunti di Dimensione

$$\frac{N}{M} : B_1, \dots, B_M$$

$$\gamma_0 = w^k$$

- FOR  $i=1, \dots, M$ :

$$Y_i = Y_{i-1} - \alpha_k \frac{1}{|B_i|} \sum_{j \in B_i} \nabla f_j(Y_{i-1})$$

END FOR

$$X^{k+1} = Y^M$$

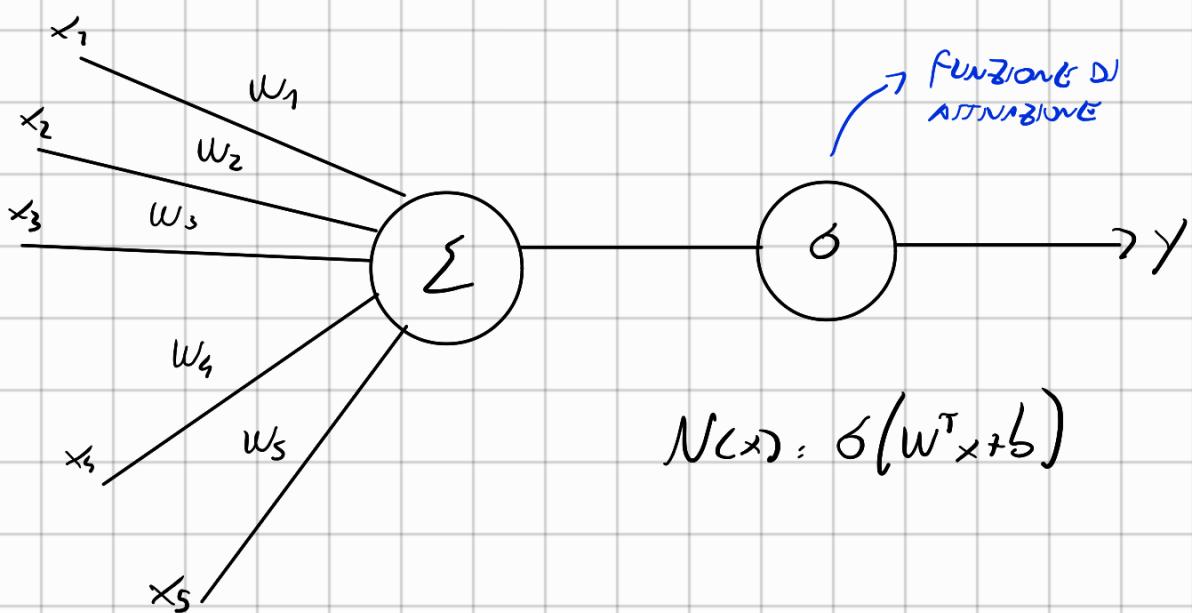
$$k = k + 1$$

END WHILE

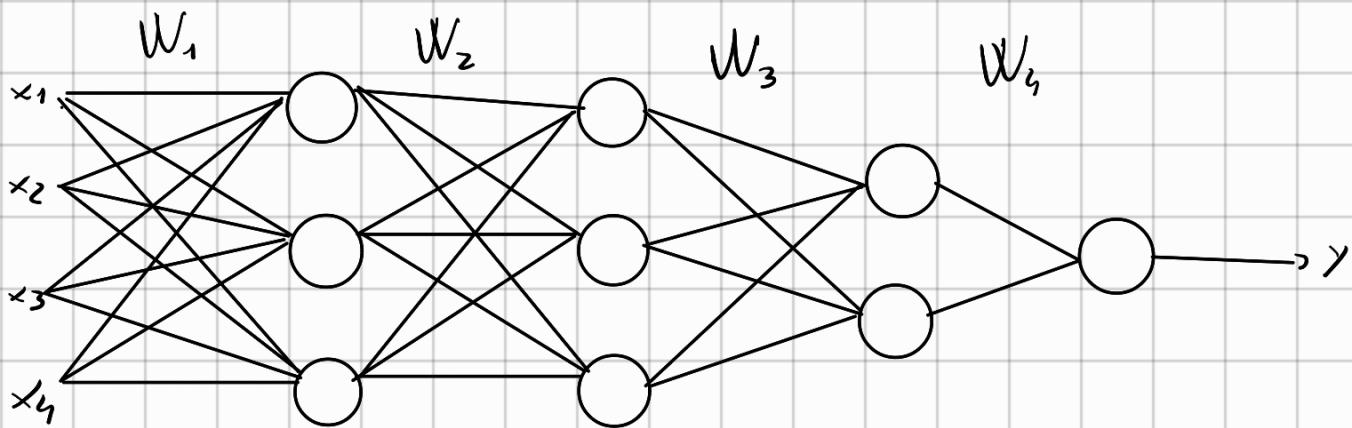
NOTIAMO CHE LE ITERAZIONI DEL CICLO WHILE SONO DELLE **EPOCHE**  
 OGNI EPOCA PREVEDE DI attraversare una volta l'intero  
 TRAINING SET.

### ADDESTRAMENTO DI RETI NEURALI

DEFINISMO IL **NEURONE ARTIFICIALE**:



NEL CASO SI ABBIANO PIÙ NEURONI:



GUARDA ALL' AUMENTARE DEL NUMERO DI LAYER AUMENTANDO SIA LA POTENZA ESpressiva DEL MODELLO, MA ANCHE LA SUA COMPLESSITÀ.

SI HA GUARDA IL SEGUENTE PROBLEMA:

$$\min_{w \in \mathbb{R}^n} \mathcal{L}(w) \sim \text{LOSS} \sim$$

- non UNARIO
- non CONVESO
- $= \sum_{i=1}^N P_i(w)$  con  $N$  grande
- $N$  GRANDE
- COSTO DI  $\nabla f(w) \propto N \times$  COSTO DI  $f(w)$

ANZUZZIAMO COSÌ COMBIN NELL' USARE **GD** VS **SGD** :

**SGD**

I DATI SONO TIPICAMENTE RIDONDANTI,  
QUINDI NON SERVE AVERE TUTTO IL  
DATASET PER OTTENERE

$$\frac{1}{|B|} \sum_{i \in B} \nabla f_i(w)$$

**GD**

• COMPLESSITÀ MINIMA:  $O(\log \frac{1}{\epsilon})$   
*(> FULL-BATCH)*

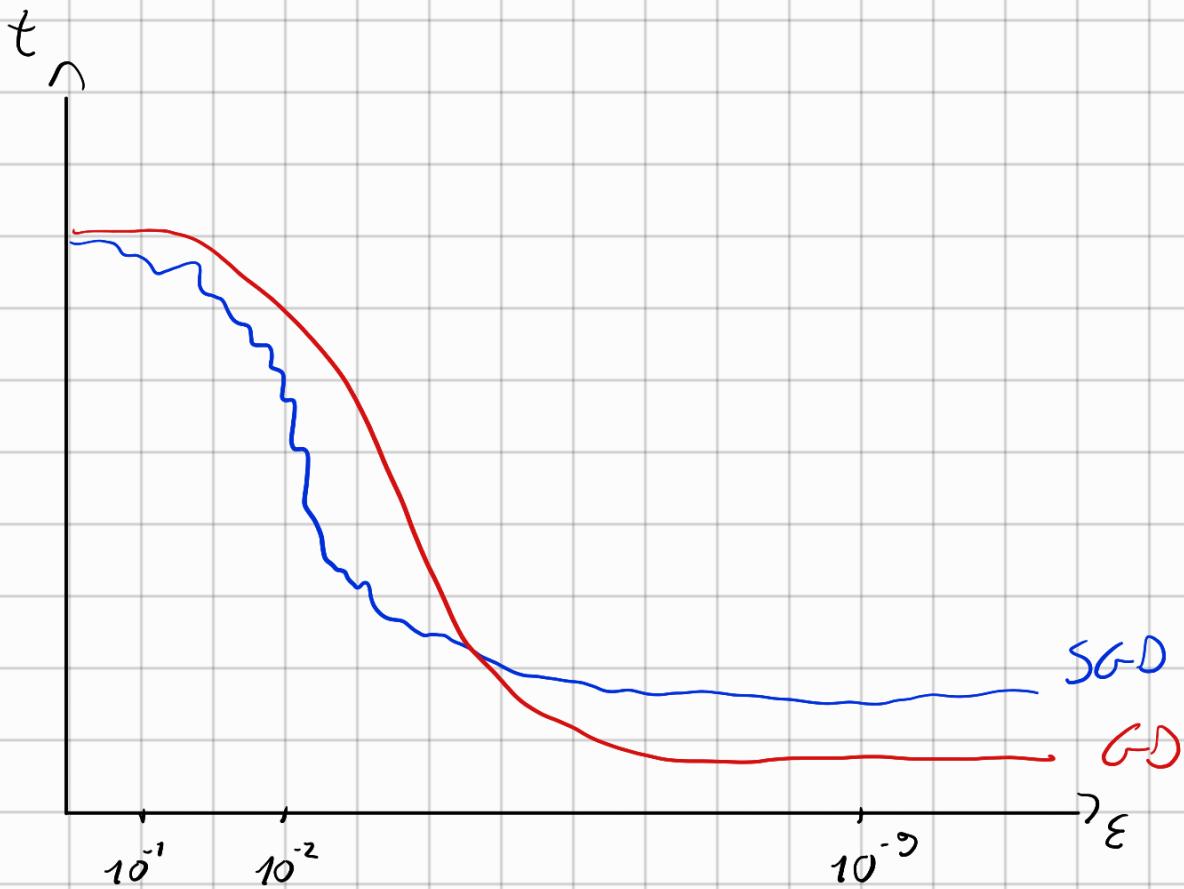
• POSSIBILITÀ DI UTILIZZARE SOLVER  
PIÙ COMPLESSI

DI DISCUSSA

• CALCOLO DI  $\mathcal{L}$  È PARALLELLIZZABILE.

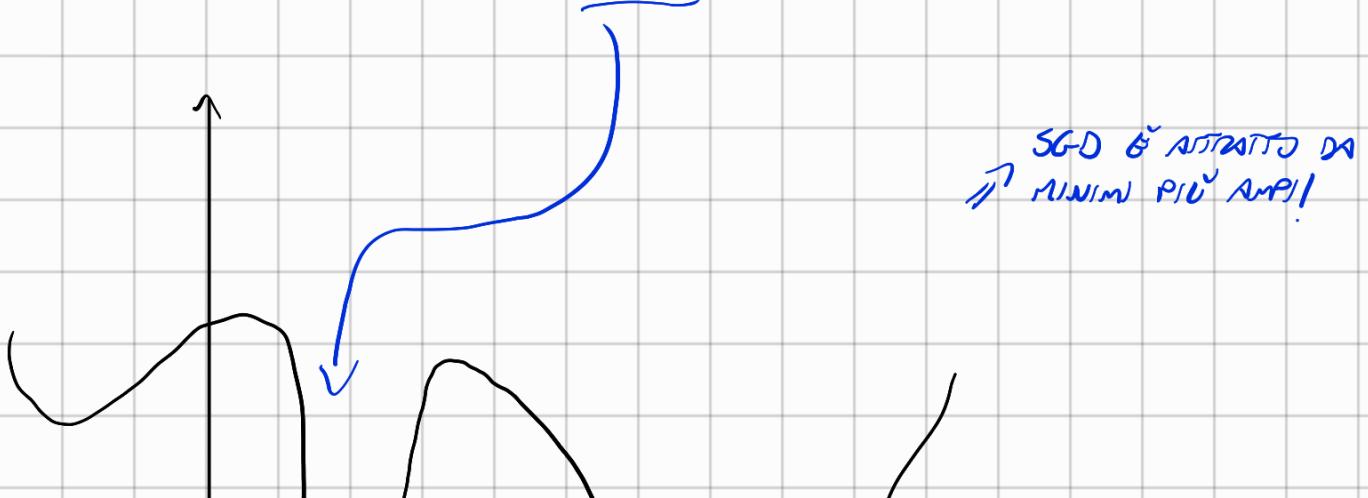
SPORADICAMENTE HA BLOCCI RIDUCITORI

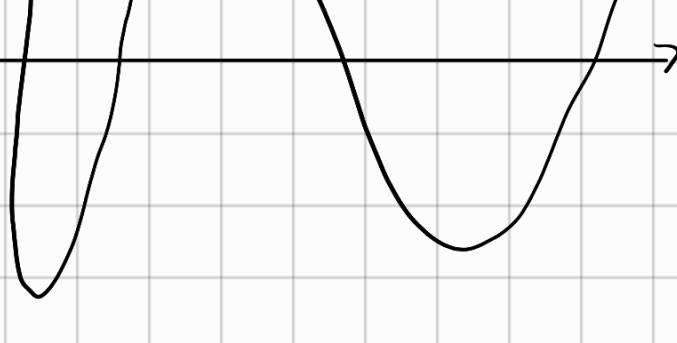
- COSTANTE  $N$  NON APPARE NELLA  
VALUTAZIONE DELLA COMPLESSITÀ



=> SOLUZIONE: SCELGO UNA VIA DI MEZZO  $\rightarrow$  MINIBATCH - SGD CON  $1 \leq B \leq n$

NOTIAMO INOLTRE CHE SGD HA UN EFFETTO INPIURO DI REGOLARIZZAZIONE, OVVERO È PIÙ DIFFICILE CADERE IN MINIMI "SHARP"





POSSIBILI MOVIMENTI PER SGD

- DIREZIONE: SI UTILIZZANO DEI TERMINI DI MOMENTUM / ACCELERAZIONE:

$$W^{k+1} = W^k - \alpha_k \nabla f_{in}(W^k) + \beta(W^k - W^{k-1})$$

IN PRATICA, SO UN PASSO, CONTRO I MOVIMENTI ANCHE DI CUI CORRENTE PARALLELA A QUELLI GIÀ CALCOLATI NEI PASSI PRECEDENTI.

UTILIZZARE IL MOMENTUM PERMETTE DI ALCUNI EVENTUALI CORRISPONDENTI MIGLIORI (ZG-ZK)

- Learning-Rate Adattivo:

$$w_i^{k+1} = w_i^k - \alpha_i^k \frac{\partial \mathcal{L}(w^k)}{\partial w_i}$$

PER L'IDEA È QUELLA DI UTILIZZARE LEARNING RATES DIVERSI AD OGNI PASSO. SI TENTA PUNTO DI STABILIRE IN MODO DA GURSICO LA DIREZIONE DEI GRADIENTI.

$\alpha_i^k$  VENGONO ALLORAWISTI AL OGNI ITERAZIONE SECONDO QUESTO REGOLA.

SCORSI DGI PCSI INIZIAU (W°)

DEVOLO ESSERE SCORSI IN modo TALE DA NON avere EXPLODING

VANISHING GRADIENTS. PER QUESTO MOTIVO SI ESEGUE LA BACKPROPAGATION

Calcolo di  $\nabla L(w)$

- DERIVATE IN FORMA SIMBOLICA  $\times$
- DIFFERENZE FINITE  $\times$
- DIFFERENZIABILITÀ AUTOMATICA ✓

=> SI SFRUTTA LA CASE RULE:  $f = f(g(x)) \Rightarrow \frac{\partial f}{\partial x_j} = \sum_i \frac{\partial f}{\partial g_i} \cdot \frac{\partial g_i}{\partial x_j}$

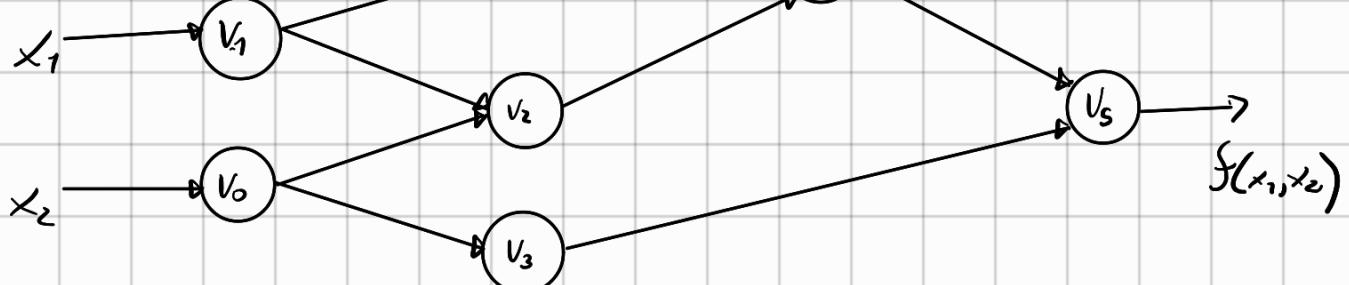
IN PARTICOLARE SI VI AD UTILIZZARE LA BACKPROPAGATION:

SUPPONIAMO DI AVERE  $f(x_1, x_2) = \log(x_1) + x_1 x_2 - \ln(x_2)$ .

NE VOLGO RICORDARE LE DERIVATI:

COSTRUISSO IL GRAFO DI COMPUTAZIONE





(2) Forward pass

$$V_{-1} = x_1$$

$$= 2$$

$$\bar{V}_{-1} = \bar{V}_1 \frac{\partial V_1}{\partial V_{-1}} + \bar{V}_2 \frac{\partial V_2}{\partial V_{-1}} = \frac{\bar{V}_1}{V_{-1}} + \bar{V}_2 V_0 = 5.5$$

$$V_0 = x_2$$

$$= 5$$

$$\bar{V}_0 = \bar{V}_3 \frac{\partial V_3}{\partial V_0} + \bar{V}_2 \frac{\partial V_2}{\partial V_0} = 1.716$$

$$V_1 = \log V_{-1}$$

$$= \log 2$$

$$\bar{V}_1 = \bar{V}_4 \frac{\partial V_4}{\partial V_1} = \bar{V}_4 = 1$$

$$V_2 = V_{-1} \cdot V_0$$

$$= 10$$

$$\bar{V}_2 = \bar{V}_4 \cdot \frac{\partial V_4}{\partial V_2} \cdot \bar{V}_3 = 1$$

$$V_3 = \ln V_0$$

$$= \ln 5$$

$$\bar{V}_3 = \bar{V}_S \frac{\partial V_S}{\partial V_3} = -\bar{V}_S = -1$$

$$V_4 = V_1 + V_2$$

$$= 0.693 + 10$$

$$\bar{V}_4 = \bar{V}_S \frac{\partial V_S}{\partial V_4} = \bar{V}_S \cdot 1 = 1$$

$$V_S = V_4 - V_3$$

$$= 10.693 + 0.693$$

$$y = \bar{V}_S = \frac{\partial Y}{\partial V_S} = \frac{\partial V_S}{\partial V_S} = 1$$

(3) Backward pass

$$y = V_S$$

$$= 11.652$$