(F94) Metodi statistici per l'apprendimento

Support Vector Machines

Docente: *Nicolò Cesa-Bianchi* versione 1 giugno 2016

La Support Vector Machine (d'ora in poi SVM) è un algoritmo di apprendimento per classificatori lineari che, fissato un training set linearmente separabile $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m) \in \mathbb{R}^d \times \{-1, +1\}$, genera il classificatore lineare corrispondente all'unica soluzione $\boldsymbol{w}^* \in \mathbb{R}^d$ del seguente problema di ottimizzazione convessa con vincoli lineari

$$\begin{array}{ll} \min\limits_{\boldsymbol{w} \in \mathbb{R}^d} & \frac{1}{2} \|\boldsymbol{w}\|^2 \\ \text{s.t.} & y_t \, \boldsymbol{w}^\top \boldsymbol{x}_t \geq 1 \quad t = 1, \ldots, m. \end{array} \tag{1}$$

Geometricamente, $\boldsymbol{w}^*$ rappresenta l'iperpiano separatore a margine massimo, come dimostrato nel seguito.

Teorema 1 *Per ogni* $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m) \in \mathbb{R}^d \times \{-1, +1\}$ *linearmente separabile, il vettore* $\boldsymbol{u}^*$ *che realizza il margine massimo*

$$\gamma^* = \max_{\boldsymbol{u} \,:\, \|\boldsymbol{u}\| = 1} \min_{t=1,\ldots,m} y_t \, \boldsymbol{u}^\top \boldsymbol{x}_t$$

soddisfa $\boldsymbol{u}^* = \gamma^* \boldsymbol{w}^*$*, dove* $\boldsymbol{w}^*$ *è soluzione di (1).*

Dimostrazione. Si noti che $\boldsymbol{u}^*$ è identificato dalla soluzione del seguente problema di ottimizzazione

$$\begin{array}{ll} \max\limits_{\gamma > 0} & \gamma^2 \\ \text{s.t.} & \|\boldsymbol{u}\|^2 = 1 \\ & y_t \, \boldsymbol{u}^\top \boldsymbol{x}_t \geq \gamma \quad t = 1, \ldots, m. \end{array}$$

Infatti, $\boldsymbol{u}$ che massimizza γ è lo stesso $\boldsymbol{u}$ che massimizza γ^2 dato che la funzione $f(\gamma) = \gamma^2$ è monotona crescente per $\gamma > 0$. Dividendo per $\gamma > 0$ entrambi i membri di ciascun vincolo $y_t \, \boldsymbol{u}^\top \boldsymbol{x}_t \geq \gamma$ otteniamo il vincolo equivalente $y_t \big(\boldsymbol{u}^\top \boldsymbol{x}_t\big) / \gamma \geq 1$. Eseguendo il cambio di variabile $\boldsymbol{w} = \boldsymbol{u}/\gamma$, e notando che $\|\boldsymbol{w}\|^2 = 1/\gamma^2$ a causa del vincolo $\|\boldsymbol{u}\|^2 = 1$, otteniamo quindi il problema equivalente

$$\begin{array}{ll} \min\limits_{\boldsymbol{w} \in \mathbb{R}^d} & \|\boldsymbol{w}\|^2 \\ \text{s.t.} & \gamma^2 \|\boldsymbol{w}\|^2 = 1 \\ & y_t \, \boldsymbol{w}^\top \boldsymbol{x}_t \geq 1 \quad t = 1, \ldots, m. \end{array}$$

Si noti ora che il vincolo $\gamma^2 \|\boldsymbol{w}\|^2 = 1$ è superfluo in quanto, per ogni $\boldsymbol{w} \in \mathbb{R}^d$, posso trovare $\gamma > 0$ tale che il vincolo è soddisfatto. Quindi lo possiamo eliminare. Scalando la funzione obiettivo per la costante $\frac{1}{2}$ otteniamo

$$\begin{array}{ll} \min\limits_{\boldsymbol{w} \in \mathbb{R}^d} & \frac{1}{2} \|\boldsymbol{w}\|^2 \\ \text{s.t.} & y_t \, \boldsymbol{w}^\top \boldsymbol{x}_t \geq 1 \quad t = 1, \ldots, m \end{array}$$

che conclude la dimostrazione □

Abbiamo quindi mostrato l'equivalente fra il problema di massimizzare il margine di $\boldsymbol{u}$ mantenendo la norma $\|\boldsymbol{u}\|$ costante ed il problema di minimizzare la norma $\|\boldsymbol{w}\|$ mantenendo il margine di $\boldsymbol{w}$ costante.

La seguente nozione ci aiuta a calcolare la forma della soluzione ottima $\boldsymbol{w}^*$.

Lemma 2 (Condizione di ottimalità di Fritz John) *Si consideri il problema*

$$\begin{array}{ll} \min\limits_{\boldsymbol{w}\in\mathbb{R}^d} & f(\boldsymbol{w}) \\ s.t. & g_t(\boldsymbol{w}) \leq 0 \quad t = 1, \ldots, m \end{array}$$

dove le funzioni $f, g_1, \ldots, g_m$ sono differenziabili. Se $\boldsymbol{w}_0$ è una soluzione ottima, allora esiste un vettore $\boldsymbol{\alpha} \in \mathbb{R}^m$ tale che

$$\nabla f(\boldsymbol{w}_0) + \sum_{t\in I} \alpha_t \nabla g_t(\boldsymbol{w}_0) = \mathbf{0}$$

dove $I = \{1 \leq t \leq m \,:\, g_t(\boldsymbol{w}_0) = 0\}$.

Applicando la condizione di Fritz John alla funzione obiettivo SVM, con $f(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2$ e $g_t(\boldsymbol{w}) = 1 - y_t\, \boldsymbol{w}^\top \boldsymbol{x}_t$ otteniamo che

$$\boldsymbol{w}^* - \sum_{t\in I} \alpha_t y_t\, \boldsymbol{x}_t = \mathbf{0}\ .$$

Quindi la soluzione ha forma

$$\boldsymbol{w}^* = \sum_{t\in I} \alpha_t y_t\, \boldsymbol{x}_t$$

dove I denota l'insieme di quegli esempi di training $(\boldsymbol{x}_t, y_t)$ tali che $y_t(\boldsymbol{w}^*)^\top \boldsymbol{x}_t = 1$. Questi $\boldsymbol{x}_t$ sono i cosiddetti *vettori di supporto*, ovvero quelle istanze di training sulle quali $\boldsymbol{w}^*$ ha margine esattamente pari a 1. Se levassimo dal training set tutti gli esempi tranne quelli di supporto la soluzione SVM non cambierebbe.

Passiamo ora a considerare il caso di un training set non linearmente separabile. Come dobbiamo cambiare la funzione obiettivo SVM? Un modo di farlo è il seguente,

$$\begin{array}{lll} \min\limits_{\boldsymbol{w}\in\mathbb{R}^d} & \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum\limits_{t=1}^{m} \xi_t & \\ \text{s.t.} & y_t\, \boldsymbol{w}^\top \boldsymbol{x}_t \geq 1 - \xi_t & t = 1, \ldots, m \\ & \xi_t \geq 0 & t = 1, \ldots, m. \end{array}$$

Le quantità ξ_t vengono dette *variabili di slack* e misurano di quanto ciascun vincolo di margine è violato da una potenziale soluzione $\boldsymbol{w}$. La media delle violazioni viene poi aggiunta alla funzione obiettivo. Un coefficiente di regolarizzazione $\lambda > 0$ è introdotto per bilanciare i due termini della funzione obiettivo.

Consideriamo ora i vincoli che coinvolgono le ξ_t, ovvero $\xi_t \geq 1 - y_t\, \boldsymbol{w}^\top \boldsymbol{x}_t$ e $\xi_t \geq 0$. Per minimizzare ciascun ξ_t, possiamo porre

$$\xi_t = \begin{cases} 1 - y_t\, \boldsymbol{w}^\top \boldsymbol{x}_t & \text{se } y_t\, \boldsymbol{w}^\top \boldsymbol{x}_t < 1 \\ 0 & \text{altrimenti.} \end{cases}$$

Ovvero, se il vincolo $y_t \, \boldsymbol{w}^\top \boldsymbol{x}_t \geq 1$ è soddisfatto da $\boldsymbol{w}$, allora ξ_t non serve e la poniamo a zero. Altrimenti, se il vincolo non è soddisfatto da $\boldsymbol{w}$, allora scegliamo il minimo valore di ξ_t che lo soddisfa, cioè $1 - y_t \, \boldsymbol{w}^\top \boldsymbol{x}_t$. Riassumendo, $\xi_t = \big[1 - y_t \, \boldsymbol{w}^\top \boldsymbol{x}_t\big]_+$, che corrisponde alla definizione hinge loss.

Ponendo $h_t(\boldsymbol{w}) = \big[1 - y_t \, \boldsymbol{w}^\top \boldsymbol{x}_t\big]_+$, il problema di SVM nel caso non linearmente separabile può allora essere riscritto come $\min_{\boldsymbol{w} \in \mathbb{R}^d} F(\boldsymbol{w})$, dove

$$F(\boldsymbol{w}) = \frac{1}{m} \sum_{t=1}^{m} h_t(\boldsymbol{w}) + \frac{\lambda}{2} \, \|\boldsymbol{w}\|^2 \ .$$

Consideriamo ora un kernel K, che implementa il prodotto interno $K(\boldsymbol{x}, \boldsymbol{x}') = \big\langle \phi_K(\boldsymbol{x}), \phi_K(\boldsymbol{x}') \big\rangle$ di due vettori $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ secondo una feature map $\phi_K : \mathbb{R}^d \to \mathcal{H}_K$. Nello spazio $\mathcal{H}_K$, la funzione $F : \mathbb{R}^d \to \mathbb{R}$ diventa $F_K : \mathcal{H}_K \to \mathbb{R}$ definita da

$$F_K(\boldsymbol{w}) = \frac{1}{m} \sum_{t=1}^{m} h_t(\boldsymbol{w}) + \frac{\lambda}{2} \, \|\boldsymbol{w}\|_K^2 \ .$$

dove $h_t(\boldsymbol{w}) = \big[1 - y_t \big\langle \boldsymbol{w}, \phi(\boldsymbol{x}_t) \big\rangle\big]_+$ e $\|\boldsymbol{w}\|_K^2 = \big\langle \boldsymbol{w}, \boldsymbol{w} \big\rangle$. Dimostriamo ora che, anche nel caso di training set non linearmente separabili in spazi kernel, soluzione $\boldsymbol{w}^* \in \mathcal{H}_K$ appartiene al sottospazio descritto dall'immagine del training set secondo la feature map ϕ_K.

Teorema 3 *Il minimo di F_K è rappresentabile come combinazione lineare di $\phi_K(\boldsymbol{x}_1), \ldots, \phi_K(\boldsymbol{x}_m)$.*

Dimostrazione. Sia $\boldsymbol{w}^*$ il minimo di F_K. Per assurdo, supponiamo che

$$\boldsymbol{w}^* = \sum_{t=1}^{m} \alpha_t \, y_t \, \phi(\boldsymbol{x}_t) + \boldsymbol{u}$$

dove $\boldsymbol{u} \in \mathcal{H}_K$ è la componente di $\boldsymbol{w}^*$ ortogonale al sottospazio descritto da $\boldsymbol{\phi}_K(\boldsymbol{x}_1), \ldots, \boldsymbol{\phi}_K(\boldsymbol{x}_m)$ (dato che $\mathcal{H}_K$ è uno spazio a prodotto interno, è sempre possibile scrivere un qualunque suo vettore in questo modo). Quindi, in particolare,

$$\big\langle \boldsymbol{u}, \phi(\boldsymbol{x}_t) \big\rangle = 0 \qquad t = 1, \ldots, m. \tag{2}$$

Ora sia $\boldsymbol{v} = \boldsymbol{w}^* - \boldsymbol{u}$. Primo, $\|\boldsymbol{v}\|_{\mathcal{V}}^2 \leq \|\boldsymbol{w}^*\|_{\mathcal{V}}^2$ dato che abbiamo tolto a $\boldsymbol{w}^*$ una componente ortogonale a $\boldsymbol{v}$, e quindi la sua lunghezza è diminuita. Secondo,

$$\begin{aligned} h_t(\boldsymbol{v}) &= \big[1 - y_t \big\langle \boldsymbol{v}, \phi(\boldsymbol{x}_t) \big\rangle\big]_+ = \big[1 - y_t \big\langle \boldsymbol{w}^* - \boldsymbol{u}, \phi(\boldsymbol{x}_t) \big\rangle\big]_+ = \big[1 - y_t \big\langle \boldsymbol{w}^*, \phi(\boldsymbol{x}_t) \big\rangle + y_t \big\langle \boldsymbol{u}, \phi(\boldsymbol{x}_t) \big\rangle\big]_+ \\ &= h_t(\boldsymbol{w}^*) \end{aligned}$$

usando (2). Quindi $F_K(\boldsymbol{v}) \leq F_K(\boldsymbol{w}^*)$, che contraddice l'ottimalità di $\boldsymbol{w}^*$. Di conseguenza, $\boldsymbol{u} = \boldsymbol{0}$. □

Notiamo che, come nel caso linearmente separabile, anche in questo caso più generale $\boldsymbol{w}^*$ dipenderà da un sottoinsieme di vettori di supporto. Ovvero, $\alpha_t \neq 0$ solo per alcuni t. A differenza del caso

lineare, questi vettori di supporto non saranno soltanto i punti del training set a margine minimo rispetto a $\boldsymbol{w}^*$, ma tutti i punti corrispondenti a variabili slack $\xi_t > 0$.

Dato che $\boldsymbol{w}^*$ è determinato da un sottinsieme dei punti del training set, possiamo limitare il rischio statistico del classificatore lineare $h^*(\boldsymbol{x}) = (\boldsymbol{w}^*)^\top \boldsymbol{x}$ usando le tecniche dei "Compression bounds". In particolare,

$$\mathrm{er}(h^*) \leq \widetilde{\mathrm{er}}(h^*) + \sqrt{\frac{1}{m}\left(N + (N+1)\ln m + \ln\frac{1}{\delta}\right)}$$

con probabilità almeno $1-\delta$ rispetto all'estrazione del training set, dove N è il numero dei vettori di supporto di $\boldsymbol{w}^*$ e $\widetilde{\mathrm{er}}(h^*)$ è la frazione degli esempi che non sono supporti e che sono classificati scorrettamente da h^*. Questo maggiorante vale immutato anche nel caso in cui l'algoritmo SVM venga eseguito in uno spazio di kernel. In tal caso, ci aspettiamo che il numero di supporti sia ridotto rispetto al caso senza kernel.

Proseguiamo mostrando come minimizzare F usando un metodo del gradiente, che poi verificheremo essere anche utilizzabile in uno spazio di kernel.

Prima di tutto osserviamo che

$$F(\boldsymbol{w}) = \frac{1}{m}\sum_{t=1}^{m} \ell_t(\boldsymbol{w})$$

dove $\ell_t(\boldsymbol{w}) = h_t(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$ è una funzione λ-fortemente convessa. Infatti, $\frac{\lambda}{2}\|\boldsymbol{w}\|^2$ è λ-fortemente convessa e h_t è convessa, il che implica che la loro somma è λ-fortemente convessa. Possiamo quindi applicare l'algoritmo OGD per funzioni fortemente convesse alle funzioni $\ell_1, \ldots, \ell_m$. Questa particolare istanza di OGD prende il nome di Pegasos e può essere descritta come segue.

Parametri: numero T di cicli, coefficiente λ di regolarizzazione
Input: Training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m) \in \mathbb{R}^d \times \{-1, +1\}$
Inizializza $\boldsymbol{w}_1 = \boldsymbol{0}$
Per $t = 1, \ldots, T$

1. Estrai uniformemente a caso un elemento $(\boldsymbol{x}_{Z_t}, y_{Z_t})$ del training set
2. $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \nabla \ell_{Z_t}(\boldsymbol{w}_t)$

Output: $\bar{\boldsymbol{w}} = \frac{1}{T}\big(\boldsymbol{w}_1 + \cdots + \boldsymbol{w}_T\big)$.

Procediamo quindi ad analizzare Pegasos. Sia $(\boldsymbol{x}_{Z_1}, y_{Z_1}), \ldots, (\boldsymbol{x}_{Z_T}, y_{Z_T})$ la sequenza di esempi del training set che sono stati estratti nel passo 1 dell'algoritmo, e sia $\ell_{Z_1}, \ldots, \ell_{Z_T}$ la corrispondente sequenza di funzioni di perdita. Cioè, $\ell_{Z_t}(\boldsymbol{w}) = h_{Z_t}(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$ dove $h_{Z_t}(\boldsymbol{w}) = \big[1 - y_{Z_t}\, \boldsymbol{w}^\top \boldsymbol{x}_{Z_t}\big]_+$.

Sia $\boldsymbol{w}^*$ la soluzione della funzione obiettivo SVM,

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \left(\frac{1}{m}\sum_{t=1}^{m} h_t(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2\right) . \tag{3}$$

Per ogni realizzazione $s_1, \ldots, s_T$ delle variabili casuali $Z_1, \ldots, Z_T$, l'analisi di OGD per funzioni fortemente convesse dà immediatamente il risultato

$$\frac{1}{T}\sum_{t=1}^{T} \ell_{s_t}(\boldsymbol{w}_t) \le \frac{1}{T}\sum_{t=1}^{T} \ell_{s_t}(\boldsymbol{w}^*) + \frac{G^2}{2\lambda T}\ln(T+1) \tag{4}$$

dove $G = \max_{t=1,\ldots,T} \|\nabla \ell_{s_t}(\boldsymbol{w}_t)\|$ è anch'essa una variabile casuale.

Per mostrare come questo risultato possa essere usato per maggiorare $F(\bar{\boldsymbol{w}})$ useremo il fatto seguente

$$\mathbb{E}\big[\ell_{Z_t}(\boldsymbol{w}_t) \mid Z_1, \ldots, Z_{t-1}\big] = \frac{1}{m}\sum_{s=1}^{m} \ell_s(\boldsymbol{w}_t) = F(\boldsymbol{w}_t)\ . \tag{5}$$

Ovvero, condizionato sulle prime $t-1$ estrazioni (le quali determinano $\boldsymbol{w}_t$), il valore atteso di $\ell_{Z_t}(\boldsymbol{w}_t)$ è pari a $F(\boldsymbol{w}_t)$. L'altro fatto che useremo è che per ogni coppia di variabili casuali X, Y vale $\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big]$. Quindi possiamo scrivere

$$\begin{aligned}
\mathbb{E}\big[F(\bar{\boldsymbol{w}})\big] &= \mathbb{E}\left[F\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t\right)\right] \\
&\le \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} F(\boldsymbol{w}_t)\right] \quad \text{usando la dis. di Jensen, dato che } F \text{ è convessa} \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[\ell_{Z_t}(\boldsymbol{w}_t) \mid Z_1, \ldots, Z_{t-1}\big]\right] \quad \text{usando (5)} \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \ell_{Z_t}(\boldsymbol{w}_t)\right] \quad \text{usando } \mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big] \\
&\le \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \ell_{Z_t}(\boldsymbol{w}^*)\right] + \frac{\mathbb{E}\big[G^2\big]}{2\lambda T}\big(\ln T + 1\big) \quad \text{usando (4)} \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[\ell_{Z_t}(\boldsymbol{w}^*) \mid Z_1, \ldots, Z_{t-1}\big]\right] + \frac{\mathbb{E}\big[G^2\big]}{2\lambda T}\big(\ln T + 1\big) \quad \text{usando } \mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big] \\
&= F(\boldsymbol{w}^*) + \frac{\mathbb{E}\big[G^2\big]}{2\lambda T}\ln(T+1) \quad \text{usando (5).}
\end{aligned}$$

Abbiamo così ottenuto

$$\mathbb{E}\big[F(\bar{\boldsymbol{w}})\big] \le F(\boldsymbol{w}^*) + \frac{\mathbb{E}\big[G^2\big]}{2\lambda T}\big(\ln T + 1\big)\ . \tag{6}$$

Quindi, se $\mathbb{E}\big[G^2\big]$ è limitato da una costante, la media $\bar{\boldsymbol{w}}$ dei vettori generati da OGD converge (in valore atteso rispetto all'estrazione a caso dei T esempi dal training set) a $\boldsymbol{w}^*$ con tasso $\frac{\ln T}{T}$. Con un po' di fatica in più è possibile dimostrare (ma non lo facciamo qui) che $\bar{\boldsymbol{w}}$ converge a $\boldsymbol{w}^*$ non solo in valore atteso ma anche in probabilità.

Ora maggioriamo il valore di G per ogni realizzazione $s_1, \ldots, s_T$ delle variabili casuali $Z_1, \ldots, Z_T$. Abbiamo $\nabla \ell_{s_t}(\boldsymbol{w}_t) = -y_{s_t}\,\boldsymbol{x}_{s_t}\,\mathbb{I}\{h_{s_t}(\boldsymbol{w}_t) > 0\} + \lambda\,\boldsymbol{w}_t$. Sia $\boldsymbol{v}_t = y_{s_t}\,\boldsymbol{x}_{s_t}\,\mathbb{I}\{h_{s_t}(\boldsymbol{w}_t) > 0\}$. Dato che

$\eta_t = 1/(\lambda t)$, notiamo che l'aggiornamento di $\boldsymbol{w}_t$ ha una forma particolarmente semplice,

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \nabla \ell_t(\boldsymbol{w}_t) = \boldsymbol{w}_t + \eta_t \boldsymbol{v}_t - \eta_t \lambda \boldsymbol{w}_t = \left(1 - \frac{1}{t}\right) \boldsymbol{w}_t + \frac{1}{\lambda t} \boldsymbol{v}_t \ .$$

Sia $X = \max_{s=1,\dots,m} \|\boldsymbol{x}_s\|$. Dato che $\|\nabla \ell_{s_t}(\boldsymbol{w}_t)\| \leq \|\boldsymbol{v}_t\| + \lambda \|\boldsymbol{w}_t\| \leq X + \lambda \|\boldsymbol{w}_t\|$, dobbiamo calcolare un maggiorante di $\|\boldsymbol{w}_t\|$. Per far ciò, esaminiamo la ricorrenza

$$\boldsymbol{w}_{t+1} = \left(1 - \frac{1}{t}\right) \boldsymbol{w}_t + \frac{1}{\lambda t} \boldsymbol{v}_t \ .$$

Come è facilmente dimostrabile per induzione, $\boldsymbol{w}_{t+1}$ è esprimibile come una combinazione lineare di $\boldsymbol{v}_s$ per $s = 1, \dots, t$ ma con quali coefficienti? Fissiamo un $s \leq t$ e notiamo che quando $\boldsymbol{v}_s$ è aggiunto a questa somma esso ha coefficiente $1/(\lambda s)$. Quando viene calcolato $\boldsymbol{w}_{t+1}$, $\boldsymbol{v}_s$ avrà coefficiente pari a

$$\frac{1}{\lambda s} \prod_{r=s+1}^{t} \left(1 - \frac{1}{r}\right) = \frac{1}{\lambda s} \prod_{r=s+1}^{t} \frac{r-1}{r} = \frac{1}{\lambda t} \ .$$

Quindi otteniamo una semplice espressione per $\boldsymbol{w}_{t+1}$,

$$\boldsymbol{w}_{t+1} = \frac{1}{\lambda t} \sum_{s=1}^{t} \boldsymbol{v}_s \ .$$

Dato che $\boldsymbol{w}_{t+1}$ è una media dei $\boldsymbol{v}_s$ divisi per λ, abbiamo infine $\|\boldsymbol{w}_{t+1}\| \leq \frac{1}{\lambda} \max_s \|\boldsymbol{v}_s\| \leq \frac{1}{\lambda} X$. Questo ci fa concludere che

$$\|\nabla \ell_t(\boldsymbol{w}_t)\| \leq X + \lambda \|\boldsymbol{w}_t\| \leq 2X \ .$$

Riportando questo maggiorante di G in (6) otteniamo

$$\mathbb{E}\big[F(\bar{\boldsymbol{w}})\big] \leq F(\boldsymbol{w}^*) + \frac{2X^2}{\lambda T} \ln(T+1) \ .$$