(F94) Metodi statistici per l'apprendimento

Online Gradient Descent

Docente: *Nicolò Cesa-Bianchi* versione 30 aprile 2018

L'analisi del Perceptrone ha rivelato come sia possibile ottenere dei maggioranti sul numero di errori di classificazione commessi dal Perceptrone su una qualsiasi sequenza di dati linearmente separabili. Qual è il significato di questo risultato? Nel modello di apprendimento statistico, dove i dati sono generati da un modello probabilistico, il criterio di valutazione di un predittore è il suo rischio statistico. Ma come valutare un classificatore quando i dati sono una sequenza qualsiasi? Il modello di apprendimento *online*, che è quello al cui interno abbiamo implicitamente analizzato il Perceptrone, suggerisce il seguente protocollo: dato un algoritmo di apprendimento A per classificazione binaria e data una sequenza arbitraria $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$ di dati.

L'algoritmo genera un modello di partenza $\boldsymbol{w}_1$
Per $t = 1, 2, \ldots$

1. Il modello corrente $\boldsymbol{w}_t$ viene testato sul prossimo esempio $(\boldsymbol{x}_t, y_t)$
2. L'algoritmo A aggiorna il modello $\boldsymbol{w}_t$ generando un nuovo modello $\boldsymbol{w}_{t+1}$

In questo protocollo di predizione sequenziale, l'algoritmo genera una sequenza $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots$ di modelli. Le prestazioni vengono valuatate misurando il rischio sequenziale, ovvero la quantità

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\{y_t \, \boldsymbol{w}_t^\top \boldsymbol{x}_t \leq 0\}$$

che conta, al variare di T, la frazione di errori di classificazione compiuta dalla sequenza di modelli sui primi T esempi. Il rischio sequenziale sostituisce la nozione di rischio statistico. Come nell'apprendimento statistico siamo interessati a studiare quanto velocemente decresce il rischio all'aumentare della taglia del training set, così nell'apprendimento online siamo interessati a studiare quanto velocemente decresce il rischio sequenziale all'aumentare di T.

Più in generale, possiamo considerare un generico problema di predizione lineare (classificazione o regressione) con funzione di perdita ℓ. Definiamo la perdita del modello $\boldsymbol{w}$ sull'esempio $(\boldsymbol{x}_t, y_t)$ come $\ell_t(\boldsymbol{w}) = \ell\big(\boldsymbol{w}^\top \boldsymbol{x}_t, y_t\big)$. Per esempio $\ell_t(\boldsymbol{w}) = \mathbb{I}\{y_t \, \boldsymbol{w}_t^\top \boldsymbol{x}_t \leq 0\}$ in classificazione, con $y_t \in \{-1, +1\}$, oppure $\ell_t(\boldsymbol{w}_t) = \big(\boldsymbol{w}^\top \boldsymbol{x}_t - y_t\big)^2$ in regressione, con $y_t \in \mathbb{R}$. In questo caso più generale valutiamo l'algoritmo di predizione tramite il rischio sequenziale,

$$\frac{1}{T} \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t)$$

dove $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots$ è la sequenza di modelli generata all'algoritmo che lavora nel protocollo di predizione sequenziale.

Il modello di apprendimento sequenziale si distingue da quello statistico perché nel primo gli algoritmi apprendono in modo incrementale, ovvero tramite ottimizzazioni progressive di un modello predittivo iniziale. Queste ottimizzazioni sono locali, cioè definite rispetto a singoli esempi della sequenza osservata. Al contrario, gli algoritmi sviluppati all'interno del modello statistico —come ad esempio il minimizzatore del rischio empirico (ERM)— operano tipicamente risolvendo un problema di ottimizzazione globale, cioè definito sull'intero training set. Il modello sequenziale è vantaggioso rispetto a quello statistico in tutte quelle situazioni dove non è possibile, oppure non è pratico, apprendere tramite ottimizzazione globale. Due esempi di situazioni del genere sono i seguenti.

1. I dati sono naturalmente generati in modo sequenziale (per esempio, dati metereologici o finanziari). Un algoritmo online che apprende incrementalmente può facilmente aggiornare un modello predittivo mano a mano che nuovi dati si rendono disponibili. Un algoritmo non incrementale, invece, deve essere riaddestrato da zero ogni volti che nuovi dati si rendono disponibili.

2. Abbiamo un training set di grandi dimensioni e perciò possiamo utilizzare soltanto algoritmi che apprendono in tempo lineare nel numero di esempi di training. Gli algoritmi online hanno esattamente questa caratteristica in quanto elaborano ciascun esempio in tempo costante. Inoltre, a differenza della maggior parte degli algoritmi che effettuano un'ottimizzazione globale, il modello corrente mantenuto dagli algoritmi sequenziali è un modello predittivo globalmente valido. Di conseguenza l'algoritmo sequenziale può essere arrestato in qualunque momento senza comprometterne l'elaborazione.

Introduciamo ora l'algoritmo sequenziale di discesa del gradiente, o *online gradient descent* (OGD). Questo algoritmo è in grado di lavorare con una qualunque funzione di perdita convessa ℓ. Per introdurre OGD, ricordiamo che una semplice tecnica per minimizzare una funzione convessa e differenziabile $\ell : \mathbb{R}^d \to \mathbb{R}$ è la discesa del gradiente. A partire da un punto arbitrario $\boldsymbol{w}_1$, la discesa del gradiente applica ripetutamente la seguente operazione: $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \nabla \ell(\boldsymbol{w}_t)$, dove $\eta_t > 0$ è un parametro. Se il punto corrente $\boldsymbol{w}_t$ non è un minimo della funzione, allora $\nabla \ell(\boldsymbol{w}_t) > 0$ e quindi $\boldsymbol{w}_{t+1}$ si sposterà in direzione del minimo della funzione. La teoria dell'ottimizzazione convessa spiega quanto velocemente la discesa del gradiente minimizza una funzione convessa rispetto al grado di convessità della funzione stessa. Per analizzare OGD, dobbiamo studiare la discesa del gradiente nel caso in cui la funzione ℓ da minimizzare cambi ad ogni passo, con una sequenza $\ell_1, \ell_2, \ldots$ ignota a priori.

Ecco una descrizione dell'algoritmo OGD con proiezione. Qui e nel seguito, assumiamo che $\ell_1, \ell_2, \ldots$ sia una sequenza di funzioni di perdita convesse e due volte differenziabili.

Algoritmo OGD con proiezione
Parametri: costante η, raggio $U > 0$
Inizializzazione: $\boldsymbol{w}_1 = \mathbf{0}$

Per $t = 1, 2, \ldots$

1. $\boldsymbol{w}'_{t+1} = \boldsymbol{w}_t - \dfrac{\eta}{\sqrt{t}} \nabla \ell_t(\boldsymbol{w}_t)$
2. $\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w} \,:\, \|\boldsymbol{w}\| \le U}{\operatorname{argmin}} \left\| \boldsymbol{w} - \boldsymbol{w}'_{t+1} \right\|$

Nel passo 2, proiettiamo $\boldsymbol{w}'_{t+1}$ in una sfera Euclidea di raggio U. Se $\left\|\boldsymbol{w}'_{t+1}\right\| \le U$, allora $\boldsymbol{w}_{t+1} = \boldsymbol{w}'_{t+1}$. Sia $\eta_t = \eta/\sqrt{t}$, dove η è un parametro dell'algoritmo.

Scopo dell'analisi è limitare la differenza fra il rischio sequenziale dell'algoritmo e quello di un qualsiasi modello $\boldsymbol{u}$ tale che $\|\boldsymbol{u}\| \le U$. Ovvero, vogliamo controllare la differenza

$$\frac{1}{T} \sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \Big) \ .$$

Più in generale, siamo interessati a dimostrare il risultato seguente. Sia

$$\boldsymbol{u}_T^* = \underset{\boldsymbol{u} \,:\, \|\boldsymbol{u}\| \le U}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} \ell_t(\boldsymbol{u})$$

il miglior predittore per i primi T passi. Allora vogliamo dimostrare

$$\frac{1}{T} \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \frac{1}{T} \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_T^*) = o(1)$$

ovvero che il rischio sequenziale di OGD converge alla perdita media del predittore ottimo $\boldsymbol{u}_T^*$ per $T \to \infty$.

L'analisi dell'algoritmo utilizza il teorema seguente.

Lemma 1 (Formula di Taylor per funzioni multivariate) *Sia $f : \mathbb{R}^d \to \mathbb{R}$ una funzione due volte differenziabile. Allora, per ogni $\boldsymbol{w}, \boldsymbol{u} \in \mathbb{R}^d$ vale*

$$f(\boldsymbol{u}) = f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^\top (\boldsymbol{u} - \boldsymbol{w}) + \frac{1}{2} (\boldsymbol{u} - \boldsymbol{w})^\top \nabla^2 f(\boldsymbol{\xi}) (\boldsymbol{u} - \boldsymbol{w})$$

dove $\nabla^2 f(\boldsymbol{\xi})$ è la matrice Hessiana di f calcolata in un punto $\boldsymbol{\xi}$ sulla retta che congiunge $\boldsymbol{u}$ a $\boldsymbol{w}$.

Fissiamo quindi $\boldsymbol{u}$ arbitrario con norma limitata da U e notiamo che, ad ogni istante t, il teorema di Taylor implica

$$\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) = \nabla \ell_t(\boldsymbol{w}_t)^\top (\boldsymbol{w}_t - \boldsymbol{u}) - \frac{1}{2} (\boldsymbol{u} - \boldsymbol{w}_t)^\top \nabla^2 \ell_t(\boldsymbol{\xi}) (\boldsymbol{u} - \boldsymbol{w}_t) \le \nabla \ell_t(\boldsymbol{w}_t)^\top (\boldsymbol{w}_t - \boldsymbol{u}) \ .$$

La disuguaglianza vale perché stiamo assumendo che ℓ_t sia due volte differenziabile e convessa, il che implica che la matrice $\nabla^2 \ell_t(\boldsymbol{\xi})$ sia positiva semidefinita. Quindi $\boldsymbol{z}^\top \nabla^2 \ell_t(\boldsymbol{\xi}) \boldsymbol{z} \geq 0$ per ogni $\boldsymbol{z} \in \mathbb{R}^d$.

Possiamo quindi procedere maggiorando la quantità $\nabla \ell_t(\boldsymbol{w}_t)^\top (\boldsymbol{w}_t - \boldsymbol{u})$,

$$\begin{aligned}
\nabla \ell_t(\boldsymbol{w}_t)^\top (\boldsymbol{w}_t - \boldsymbol{u}) &= -\frac{1}{\eta_t} (\boldsymbol{w}'_{t+1} - \boldsymbol{w}_t)^\top (\boldsymbol{w}_t - \boldsymbol{u}) \\
&= \frac{1}{\eta_t} \left(\frac{1}{2} \|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \frac{1}{2} \|\boldsymbol{w}'_{t+1} - \boldsymbol{u}\|^2 + \frac{1}{2} \|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_t\|^2 \right) \\
&\leq \frac{1}{\eta_t} \left(\frac{1}{2} \|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \frac{1}{2} \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 + \frac{1}{2} \|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_t\|^2 \right) . \qquad (1)
\end{aligned}$$

La prima uguaglianza usa il fatto che $\boldsymbol{w}'_{t+1} - \boldsymbol{w}_t = \eta_t \nabla \ell_t(\boldsymbol{w}_t)$. La seconda è un'identità algebrica che si verifica rapidamente facendo i conti. Infine la disuguaglianza vale perché $\boldsymbol{u}$ appartiene alla sfera di raggio U centrata sull'origine, e quindi proiettando $\boldsymbol{w}'_{t+1}$ su questa sfera la distanza con $\boldsymbol{u}$ non può aumentare.

Ora aggiungiamo e togliamo lo stesso termine $\frac{1}{2\eta_{t+1}} \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2$ all'ultimo membro della catena di disuguaglianze mostrata sopra. Poi raggruppiamo i termini come indicato qua sotto

$$\underbrace{\frac{1}{2\eta_t} \|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \frac{1}{2\eta_{t+1}} \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2}_{} \underbrace{- \frac{1}{2\eta_t} \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 + \frac{1}{2\eta_{t+1}} \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2}_{} + \frac{1}{2\eta_t} \|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_t\|^2 .$$

Sommando su $t = 1, \ldots, T$ notiamo che i primi due termini sono una somma telescopica, mentre i secondi due termini hanno un fattore comune,

$$\begin{aligned}
\sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \Big) &\leq \frac{1}{2\eta} \|\boldsymbol{w}_1 - \boldsymbol{u}\|^2 - \frac{1}{2\eta_{T+1}} \|\boldsymbol{w}_{T+1} - \boldsymbol{u}\|^2 \\
&\quad + \frac{1}{2} \sum_{t=1}^{T} \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) + \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\eta_t} \|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_t\|^2 . \qquad (2)
\end{aligned}$$

Ora usiamo i seguenti fatti:

$$\begin{aligned}
\boldsymbol{w}_1 = \boldsymbol{0} \quad & \text{per definizione di OGD} \\
\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 \leq 4U^2 \quad & \text{dato che sia } \boldsymbol{w}_{t+1} \text{ che } \boldsymbol{u} \text{ appartengono alla sfera di raggio } U \\
\|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_t\|^2 = \eta_t^2 \|\nabla \ell_t(\boldsymbol{w}_t)\|^2 \quad & \text{per definizione di OGD.}
\end{aligned}$$

Sostituendo queste relazioni nell'ultima disuguaglianza e scegliendo G tale che $\|\nabla \ell_t(\boldsymbol{w}_t)\| \leq G$ per ogni t, otteniamo

$$\begin{aligned}
\sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \Big) &\leq \frac{U^2}{2\eta} - \frac{1}{2\eta_{T+1}} \|\boldsymbol{w}_{T+1} - \boldsymbol{u}\|^2 \\
&\quad + 2U^2 \sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) + \frac{1}{2\eta_{T+1}} \|\boldsymbol{w}_{T+1} - \boldsymbol{u}\|^2 - \frac{1}{2\eta_T} \|\boldsymbol{w}_{T+1} - \boldsymbol{u}\|^2 + \frac{G^2}{2} \sum_{t=1}^{T} \eta_t .
\end{aligned}$$

Ora semplifichiamo la somma telescopica, cancelliamo i termini con segno opposto e maggioriamo omettendo il termine $-\frac{1}{2\eta_T}\left\|\boldsymbol{w}_{T+1}-\boldsymbol{u}\right\|^2$,

$$\begin{aligned}\sum_{t=1}^{T}\Big(\ell_t(\boldsymbol{w}_t)-\ell_t(\boldsymbol{u})\Big) &\leq \frac{U^2}{2\eta}+\frac{2U^2}{\eta_T}-\frac{2U^2}{\eta}+\frac{G^2}{2}\sum_{t=1}^{T}\eta_t \leq \frac{2U^2\sqrt{T}}{\eta}+\frac{G^2\eta}{2}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \\ &\leq \frac{2U^2\sqrt{T}}{\eta}+G^2\eta\sqrt{T}\end{aligned}$$

dove abbiamo usato la maggiorazione

$$\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T}\ .$$

Scegliendo $\eta=\big(U/G\big)\sqrt{2}$ e dividendo tutto per T otteniamo il risultato finale

$$\frac{1}{T}\sum_{t=1}^{T}\ell_t(\boldsymbol{w}_t) \leq \min_{\boldsymbol{u}\,:\,\|\boldsymbol{u}\|\leq U}\frac{1}{T}\sum_{t=1}^{T}\ell_t(\boldsymbol{u})+UG\sqrt{\frac{8}{T}}\ . \tag{3}$$

È possibile ottenere un valore esplicito per G facendo assunzioni particolari. Per esempio, $\ell_t(\boldsymbol{w})=\big(\boldsymbol{w}^\top\boldsymbol{x}_t-y_t\big)^2$, ovvero regressione con funzione di perdita quadratica. Assumendo $\|\boldsymbol{x}_t\|\leq X$ e $|y_t|\leq UX$ per ogni t, possiamo calcolare

$$\|\nabla\ell_t(\boldsymbol{w}_t)\| \leq 2\big|\boldsymbol{w}^\top\boldsymbol{x}_t-y_t\big|\,\|\boldsymbol{x}_t\| \leq 2\big(\|\boldsymbol{w}_t\|\,\|\boldsymbol{x}_t\|+|y_t|\big)\,\|\boldsymbol{x}_t\| \leq 4UX^2\ .$$

Sostituendo questo valore di G nel maggiorante precedente otteniamo

$$\frac{1}{T}\sum_{t=1}^{T}\ell_t(\boldsymbol{w}_t) \leq \min_{\boldsymbol{u}\,:\,\|\boldsymbol{u}\|\leq U}\frac{1}{T}\sum_{t=1}^{T}\ell_t(\boldsymbol{u})+8(UX)^2\sqrt{\frac{2}{T}}\ .$$

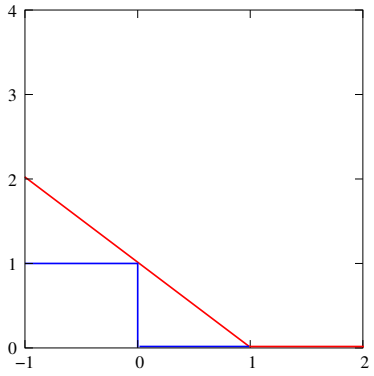

Figura 1: La hinge loss $h(z)=[1-z]_+$ (in rosso) è un maggiorante convesso alla funzione di perdita zero-uno $\ell(z)=\mathbb{I}\{z\leq 0\}$ (errore di classificazione binaria, in blu).

Notiamo ora che possiamo esprimere l'algoritmo del Perceptrone come un caso particolare di OGD. Infatti, possiamo scrivere la regola di aggiornamento del Perceptrone come discesa del gradiente su una particolare funzione di perdita chiamata hinge loss: $h_t(\boldsymbol{w}) = \left[1 - y_t\, \boldsymbol{w}^\top \boldsymbol{x}_t\right]_+$, dove $[z]_+ = \max\{0, z\}$. Questa funzione è convessa e maggiora la funzione indicatrice di errore, ovvero $\mathbb{I}\{z \le 0\} \le [1-z]_+$ per ogni $z \in \mathbb{R}$ —si veda la Figura 1. Il gradiente della hinge loss è facilmente calcolato come

$$\nabla h_t(\boldsymbol{w}) = \begin{cases} -y_t \boldsymbol{x}_t & \text{se } y_t \boldsymbol{w}^\top \boldsymbol{x}_t \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Si noti che $[1-z]_+$ non è differenziabile in $z = 1$, ma l'analisi funziona scegliendo un qualsiasi valore fra -1 e 0 come valore della derivata di $[1-z]_+$ in 1.

Per definire il Perceptrone come istanza di OGD dobbiamo aggiungere la condizione che l'aggiornamento venga fatto solo quando il modello corrente $\boldsymbol{w}_t$ sbaglia a classificare $(\boldsymbol{x}_t, y_t)$,

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \nabla h_t(\boldsymbol{w}_t) \mathbb{I}\{y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t \le 0\} = \boldsymbol{w}_t + \eta_t\, y_t \boldsymbol{x}_t \mathbb{I}\{y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t \le 0\}\ . \tag{4}$$

Dato che $\boldsymbol{w}_t$ cambia solo quando $y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t \le 0$, possiamo applicare l'analisi di OGD ai soli passi t dove $\boldsymbol{w}_t$ sbaglia, cioè ai passi t dove $y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t \le 0$. Inoltre, scegliamo $\eta_t = \eta$ per ogni t e omettiamo la proiezione di $\boldsymbol{w}'_{t+1}$ nella sfera di raggio U, cioè poniamo $\boldsymbol{w}_{t+1} = \boldsymbol{w}'_{t+1}$. La disuguaglianza (2), omettendo il termine negativo $-\frac{1}{2\eta_1} \left\|\boldsymbol{w}_{T+1} - \boldsymbol{u}\right\|^2$, ci dà

$$\begin{aligned} \sum_{t=1}^{T} \Big(h_t(\boldsymbol{w}_t) - h_t(\boldsymbol{u})\Big) \mathbb{I}\{y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t \le 0\} &\le \frac{1}{2\eta} \left\|\boldsymbol{u}\right\|^2 \\ &+ \frac{1}{2} \sum_{t=1}^{T} \left\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\right\|^2 \left(\frac{1}{\eta} - \frac{1}{\eta}\right) \mathbb{I}\{y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t \le 0\} + \frac{\eta G^2}{2} \sum_{t=1}^{T} \mathbb{I}\{y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t \le 0\} \end{aligned}$$

per un qualunque $\boldsymbol{u} \in \mathbb{R}^d$. Si noti che i termini della prima sommatoria nel membro destro della disuguaglianza sono tutti pari a zero (e questo è il motivo per cui possiamo evitare le proiezioni). Quindi, dato che $y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t \le 0$ implica $h_t(\boldsymbol{w}_t) \ge 1$, e ponendo $X = \max_t \left\|\boldsymbol{x}_t\right\| = \max_t \left\|\nabla h_t(\boldsymbol{w}_t)\right\|$ così da avere $X = G$, otteniamo

$$\sum_{t=1}^{T} \mathbb{I}\{y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t \le 0\} \le \sum_{t=1}^{T} h_t(\boldsymbol{u}) + \frac{1}{2\eta} \left\|\boldsymbol{u}\right\|^2 + \frac{\eta X^2}{2} \sum_{t=1}^{T} \mathbb{I}\{y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t \le 0\}\ .$$

Sia $M_T = \sum_{t=1}^{T} \mathbb{I}\{y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t \le 0\}$ il numero di errori compiuti dal Perceptrone nei primi T passi. Scegliendo $\eta = \left\|\boldsymbol{u}\right\| / \left(X\sqrt{M_T}\right)$, risolvendo per M_T e maggiorando otteniamo

$$M_T \le \sum_{t=1}^{T} h_t(\boldsymbol{u}) + \left(\left\|\boldsymbol{u}\right\| X\right)^2 + \left\|\boldsymbol{u}\right\| X \sqrt{\sum_{t=1}^{T} h_t(\boldsymbol{u})}\ .$$

Questo è il maggiorante al numero di errori del Perceptrone nel caso generale (sequenze non linearmente separabili). Si noti che quando la sequenza è linearmente separabile, allora esiste $\boldsymbol{u} \in \mathbb{R}^d$ tale che $y_t\, \boldsymbol{u}^\top \boldsymbol{x}_t \ge 1$ per ogni t, il che implica $h_t(\boldsymbol{u}) = 0$ per ogni t. Quindi il maggiorante si riduce a

$$M_T \le \left(\left\|\boldsymbol{u}\right\| X\right)^2$$

che corrisponde al teorema di convergenza del Perceptrone.

OGD con hinge loss ha però alcune differenze rispetto al Perceptrone. Per prima cosa, mentre nel Perceptrone $\eta = 1$, qui η dev'essere scelto in base a $\|u\|$, X e M_T. In realtà questa differenza è fittizia. Infatti, come si nota dalla regola (4) di aggiornamento, il peso $\boldsymbol{w}_t$ ha la forma

$$\boldsymbol{w}_t = \eta \sum_{s=1}^{t-1} y_s \boldsymbol{x}_s \mathbb{I}\{y_s \boldsymbol{w}_s^\top \boldsymbol{x}_s \leq 0\} \ .$$

Dato che la predizione è $\operatorname{sgn}\big(\boldsymbol{w}_t^\top \boldsymbol{x}_t\big)$, il valore di $\eta > 0$ è completamente ininfluente. In altre parole, l'algoritmo eseguito con $\eta = 1$ e quello eseguito con $\eta = \|\boldsymbol{u}\| \,/\, \big(X\sqrt{M_T}\big)$ hanno esattamente lo stesso comportamento. Quindi possiamo assumere senza perdita di generalità che l'algoritmo venga eseguito con $\eta = 1$, come il Perceptrone.

Il maggiorante (3) vale per qualunque sequenza $\ell_1, \ell_2, \ldots$ di funzioni di perdita convesse, quindi anche funzioni lineari, per esempio $\ell_t(\boldsymbol{w}) = \big|y_t - \boldsymbol{w}^\top \boldsymbol{x}_t\big|$ per $\boldsymbol{x}_t \in \mathbb{R}^d$ e $y_t \in \mathbb{R}$. È possibile dimostrare che se le funzioni di perdita sono effettivamente tutte lineari, non è possibile migliorare (3). Ma cosa succede se invece le funzioni di perdita sono convesse e mai piatte? Per definire questa situazione ricorriamo alla nozione di *convessità forte*. Una funzione differenziabile ℓ è σ-fortemente convessa, per un dato $\sigma > 0$, se

$$\ell(\boldsymbol{w}) - \ell(\boldsymbol{u}) \leq \nabla \ell(\boldsymbol{w})^\top (\boldsymbol{w} - \boldsymbol{u}) - \frac{\sigma}{2} \|\boldsymbol{u} - \boldsymbol{w}\|^2 \ .$$

Equivalentemente, possiamo dire che la matrice Hessiana di ℓ ha rango pieno, oppure che ha gli autovalori tutti strettamente maggiori di zero. Un semplice esempio di funzione fortemente convessa è $\ell(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2$. Infatti,

$$\frac{1}{2}\|\boldsymbol{w}\|^2 - \frac{1}{2}\|\boldsymbol{u}\|^2 = \boldsymbol{w}^\top(\boldsymbol{w} - \boldsymbol{u}) - \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{u}\|^2$$

Quindi la funzione è fortemente convessa per $\sigma = 1$.

L'algoritmo OGD per funzioni fortemente convesse non ha bisogno del passo di proiezione ed è quindi completamente privo di parametri.

Algoritmo OGD senza proiezione per funzioni fortemente convesse
Inizializzazione: $\boldsymbol{w}_1 = \boldsymbol{0}$
Per $t = 1, 2, \ldots$

1. $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \nabla \ell_t(\boldsymbol{w}_t)$

Per l'analisi, ripetiamo il passo (1) dell'analisi di OGD sfruttando l'assunzione che $\ell_1, \ell_2, \ldots$ sono tutte funzioni σ-fortemente convesse,

$$\begin{aligned}
\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) &\leq \nabla \ell_t(\boldsymbol{w}_t)^\top (\boldsymbol{w}_t - \boldsymbol{u}) - \frac{\sigma}{2}\|\boldsymbol{u} - \boldsymbol{w}_t\|^2 \\
&= -\frac{1}{\eta_t}(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t)^\top (\boldsymbol{w}_t - \boldsymbol{u}) - \frac{\sigma}{2}\|\boldsymbol{u} - \boldsymbol{w}_t\|^2 \\
&\leq \frac{1}{\eta_t}\left(\frac{1}{2}\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \frac{1}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 + \frac{1}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2\right) - \frac{\sigma}{2}\|\boldsymbol{u} - \boldsymbol{w}_t\|^2 \ .
\end{aligned}$$

Procedendo in modo completamente analogo al caso di OGD con proiezione, ma sfruttando la presenza dei termini aggiuntivi $-\frac{\sigma}{2}\left\|\boldsymbol{u}-\boldsymbol{w}_t\right\|^2$ otteniamo

$$\sum_{t=1}^{T}\Big(\ell_t(\boldsymbol{w}_t)-\ell_t(\boldsymbol{u})\Big) \leq \left(\frac{1}{\eta}-\sigma\right)\frac{1}{2}\left\|\boldsymbol{w}_1-\boldsymbol{u}\right\|^2 - \frac{1}{2\eta_{T+1}}\left\|\boldsymbol{w}_{T+1}-\boldsymbol{u}\right\|^2 + \frac{1}{2}\sum_{t=1}^{T-1}\left\|\boldsymbol{w}_{t+1}-\boldsymbol{u}\right\|^2\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}-\sigma\right) + \left\|\boldsymbol{w}_{T+1}-\boldsymbol{u}\right\|^2\frac{1}{2}\left(\frac{1}{\eta_{T+1}}-\frac{1}{\eta_T}\right) + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t$$

dove, analogamente a prima, $G \geq \max_t \left\|\nabla \ell_t(\boldsymbol{w}_t)\right\|$.

Omettendo il termine negativo $-\frac{1}{2\eta_T}\left\|\boldsymbol{w}_{T+1}-\boldsymbol{u}\right\|^2$, semplificando il termine $\frac{1}{2\eta_{T+1}}\left\|\boldsymbol{w}_{T+1}-\boldsymbol{u}\right\|^2$ che appare con segni opposti e utilizzando la scelta $\eta_t = \frac{1}{\sigma t}$, osserviamo alcune ulteriori sorprendenti semplificazioni che ci conducono a

$$\sum_{t=1}^{T}\Big(\ell_t(\boldsymbol{w}_t)-\ell_t(\boldsymbol{u})\Big) \leq \frac{G^2}{2\sigma}\sum_{t=1}^{T}\frac{1}{t} \leq \frac{G^2}{2\sigma}\ln(T+1)$$

dove abbiamo usato un semplice maggiorante logaritmico alla somma armonica $1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{T}$.

Questo implica il risultato finale

$$\frac{1}{T}\sum_{t=1}^{T}\ell_t(\boldsymbol{w}_t) \leq \min_{\boldsymbol{u}\in\mathbb{R}^d}\frac{1}{T}\sum_{t=1}^{T}\ell_t(\boldsymbol{u}) + \frac{G^2}{2\sigma}\frac{\ln(T+1)}{T}\ .$$

Possiamo confrontare il rischio sequenziale appena dimostrato per le funzioni di perdita fortemente convesse con quello ottenuto in (3) per le funzioni semplicemente convesse.