Online Learning and Online Convex Optimization
Shai-Shalev Shwartz

Stefanos Poulis

February 6, 2014
Outline

1. Notation-Definitions
   - Lipchitzness
   - Convexity
   - Sub-gradients

2. Follow The Leader (FTL)

3. Strongly Convex Regularizers

4. Follow The Regularized Leader (FoReL)
   - Online Gradient Descent (OGD)
   - Exponentiated Gradient Descent (EGD)

5. Online Classification
   - Perceptron
   - Winnow
Definition

A function $f$ is called $L$-Lipschitz over a set $S$ with respect to a norm $\| \ast \|$ if for all $u, w \in S$ we have $|f(u) - f(w)| \leq L\|u - w\|$. 

Furthermore, such $z$ is called the sub-gradient of $f$ at $w$. 

Stefanos Poulis (UCSD)
Notation-Definitions

**Definition**

A function \( f \) is called \( L \)-Lipschitz over a set \( S \) with respect to a norm \( \| * \| \) if for all \( u, w \in S \) we have \( |f(u) - f(w)| \leq L \| u - w \| \).

**Definition**

A set \( S \) is convex if for all \( u, w \in S \) and \( \alpha \in [0, 1] \) we have that \( \alpha u + (1 - \alpha)w \in S \) as well. A function \( f : S : \mathbb{R} \) is convex iff for all \( w \in S \) there exists \( z \) such that

\[
\forall u \in S, f(u) \geq f(w) + (u - w, z). \tag{1}
\]

Furthermore, such \( z \) is called the **sub-gradient** of \( f \) at \( w \).
Follow-The-Leader (FTL)

Algorithm: Follow-The-Leader

\[ \forall t, w_t = \arg\min_{w \in S} \sum_{i=1}^{t-1} f_i(w) \]  \hspace{1cm} (2)

Lemma

Let \( w_1, w_2, ... \) be the sequence of vectors produced by FTL. Then for all \( u \in S \) we have

\[ \text{Regret}_T(u) = \sum_{t=1}^{T} (f_t(w_t) - f_t(u)) \leq \sum_{t=1}^{T} (f_t(w_t) - f_t(w_{t+1})) \]  \hspace{1cm} (3)

Proof.

Sketch: Use induction
Follow-the-Regularized-Leader (FoReL)

Algorithm: Follow-the-Regularized-Leader

\[ \forall t, w_t = \arg\min_{w \in S} \sum_{i=1}^{t-1} f_i(w) + R(w) \]  

- \( R : S \rightarrow \mathbb{R} \) is a regularization term
- The goal of regularization is to stabilize the solution
Consider $f_t = \langle w, z \rangle$, let $S = \mathbb{R}^d$ and run FoReL with $R(w) = \frac{1}{2\eta} \| w \|_2^2$, where $\eta \geq 0$. Then, the gradient updates are

$$w_{t+1} = -\eta \sum_{i=1}^{t} z_i = w_t - \eta Z_t$$

(5)

This rule is often called Online Gradient Descent.
Follow-the-Regularized-Leader

Theorem

Consider running FoReL on a sequence of linear functions, \( f_t(w) = \langle w, z_t \rangle \) for all \( t \), with \( S = \mathbb{R}^d \) and with the regularizer \( R(w) = \frac{1}{2\eta} \| w \|_2^2 \), which yields the predictions given by the gradient-updates. Then, for all \( u \) we have,

\[
\text{Regret}_T(u) \leq \frac{1}{2\eta} \| u \|_2^2 + \eta \sum_{t=1}^{T} \| z_t \|_2^2.
\]  

(6)

Proof.

Sketch: Run FTL on \( f_0, f_1, \ldots, f_T \), where \( f_0 = R \)

Use gradient updates
Online Gradient Descent (OGD)

Running FoReL with Euclidean regularization yields OGD

Algorithm: Online Gradient Descent

- Parameter: $\eta > 0$
- Initialize: $w_1 = 0$
- Update rule: $w_{t+1} = w_t - \eta z_t$

OGD enjoys the same bound as FoReL, namely

$$\text{Regret}_T(u) \leq \frac{1}{2\eta} \|u\|_2^2 + \eta \sum_{t=1}^{T} \|z_t\|_2^2. \quad (7)$$
Better bound for OGD

**Lemma**

Let $f : S \to \mathbb{R}$ be convex. Then $f$ is $L$-Lipschitz over $S$ with respect to a norm $\| \cdot \|$ iff for all $w \in S$ and $z \in \partial f(w)$ we have that $\| z \|_* \leq L$, where $\| \cdot \|_*$ is the dual norm.

**Corollary**

Consider previous bound for OGD,

$$
\text{Regret}_T(u) \leq \frac{1}{2\eta} \| u \|^2 + \eta \sum_{t=1}^T \| z_t \|^2.
$$

(8)

If we further assume that each $f_t$ is $L_t$-Lipschitz with respect to $\| \cdot \|_2$, and let $L$ be such that $\frac{1}{T} \sum_{t=1}^T L_t^2 \leq L^2$, then

$$
\text{Regret}_T(u) \leq \frac{1}{2\eta} \| u \|^2 + \eta TL^2.
$$

(9)
A function is strongly convex if it is strictly above its tangent.

**Definition**

A function \( f : S \to \mathbb{R}^d \) is \( \sigma \)-strongly-convex over \( S \) with respect to a norm \( \| . \| \) if for any \( w \in S \) we have

\[
\forall z \in \partial f(w), \forall u \in S, f(u) \geq f(w) + \langle z, u - w \rangle + \frac{\sigma}{2} \| u - w \|^2.
\] (10)

**Example**

\[ R(w) = \frac{1}{2} \| w \|^2 \] is 1-strongly-convex with respect to the \( l_2 \) norm over \( \mathbb{R}^d \).

**Example**

\[ R(w) = \sum_{i=1}^{d} w_i \log(w_i) \] is \( \frac{1}{B} \)-strongly-convex with respect to the \( l_1 \) norm over the set \( S = \{ w \in \mathbb{R}^d : w > 0 \land \| w \|_1 \leq B \} \).
Analyzing FoReL with Strongly Convex Regularizers

**Theorem**

Let \( f(1), \ldots, f(T) \) be a sequence of convex functions such that \( f_t \) is \( L_t \)-Lipschitz with respect to some norm \( \| \cdot \| \). Let \( L \) be such that
\[
\frac{1}{T} \sum_t L_t^2 \leq L^2.
\]
Assume that FoReL is run on the sequence with a regularization function that is \( \sigma \)-strongly-convex with respect to the same norm. Then for all \( u \in S \),

\[
\text{Regret}_T(u) \leq R(u) - \min_{w \in S} R(w) + \frac{TL^2}{\sigma} \tag{11}
\]

**Proof.**

Sketch: Use the fact that \( f_t(w_t) - f_t(w_{t+1}) \leq \frac{L_t^2}{\sigma} \).
Derived Algorithms

- Running FoReL with $R(w) = \frac{1}{2}\|w\|_2^2$ yields Online Gradient Descent, with updates
  \[ w_{t+1} = w_t - \eta Z_t \]  \hspace{1cm} (12)

- Running FoReL with $R(w) = \sum_{i=1}^{d} w_i \log(w_i)$ yields Exponentiated Gradient Descent, with updates
  \[ w_{t+1}(i) = w_t(i) e^{\eta z_t(i)} \]  \hspace{1cm} (13)
Algorithm: Exponentiated Gradient Descent (Un-normalized)

- parameter: $\eta > 0$
- initialize: $w_1 = (1/d, \ldots, 1/d)$
- update rule: $\forall i, w_{t+1}(i) = w_t(i)e^{-\eta z_t(i)}$

Theorem

Let $f(1), \ldots, f(T)$ be a sequence of convex functions such that $f_t$ is $L_t$-Lipschitz with respect to some norm $\|\cdot\|$. Let $L$ be such that $\frac{1}{T}\sum_t L_t^2 \leq L^2$. Assume Exponentiated Gradient Descent is run on the sequence and with the set $S = \{w : \|w\|_1 = B \land w > 0\} \subset \mathbb{R}^d$. Then,

$$\text{Regret}_T(S) \leq \frac{B \log(d)}{\eta} + \eta BTL^2.$$  \hspace{1cm} (14)

Proof.

Sketch: Use strong convexity and Holder’s inequality.
Online Classification
- $y \in \{-1, 1\}$
- A weight vector $w$ makes a mistake on an example $(x, y)$ whenever $\text{sign}(\langle w, x \rangle) \neq y$
- 0-1 loss $l(w, (x, y)) = I[y\langle w, x \rangle \leq 0]$
- Define surrogate loss $f_t = [1 - y\langle w, x \rangle]_+$, (hinge-loss)
- $f_t$ is convex and for all $w$, $f_t(w) \geq 0$-1 loss
Run Online Gradient Descent on the sequence of functions $f_t(w)$ using update rule $w_{t+1} = w_t - \eta z_t$, where $z_t \in \partial f_t(w)$. We can check that $z_t = -y_t x_t \in \partial f_t(w)$.

Obtain update rule

$$w_{t+1} = \begin{cases} w_t, & y_t(w_t, x_t) > 0 \\ w_t + \eta y_t x_t, & otherwise \end{cases}$$
Algorithm: Perceptron

initialize: \( w_1 = 0 \)

for \( t = 1, 2, \ldots, T \)
    receive \( x_t \)
    predict \( p_t = \text{sign}(\langle w_t, x_t \rangle) \)
    if \( y_t(\langle w_t, x_t \rangle) \leq 0 \)
        \( w_{t+1} = w_t + y_t x_t \)
    else \( w_{t+1} = w_t \)
Theorem

Suppose that the Perceptron runs on a sequence \((x_1, y_1, \ldots, x_T, y_T)\) and let \(R = \|x_t\|_\infty\). Let \(M\) be the rounds on which the Perceptron errs and let 
\[
f_t(w) = l_{i \in M}[1 - y_t \langle w, x_t \rangle] +
\]
\[
M \leq \sum_t f_t(u) + R\|u\| \left(\sum_t f_t(u)\right)^{\frac{1}{2}} + R^2\|u\|^2
\]

(15)
Perceptron

**Theorem**

Suppose that the Perceptron runs on a sequence \((x_1, y_1, \ldots, x_T, y_T)\) and let \(R = \|x_t\|_{\infty}\). Let \(M\) be the rounds on which the Perceptron errs and let 
\[
f_t(w) = l_{[i \in M]}[1 - y_t \langle w, x_t \rangle] +
\]

\[
M \leq \sum_t f_t(u) + R \|u\| (\sum_t f_t(u))^{\frac{1}{2}} + R^2 \|u\|^2
\]  

(15)

**Proof.**

Sketch: Follow analysis for OGD and use claim that given \(x, b, c \in \mathbb{R}^+\), \(x \leq c + b^2 + bc^{1/2}\)
Winnow

- $y \in \{-1, 1\}$
- Originally proposed for the class of $k$ monotone Boolean functions
- $\langle w, x \rangle \geq 1$, if one of the relevant features is turned on in $x$. Otherwise, $\langle w, x \rangle = 0$
- A weight vector $w$ errs on $(x, y)$ if $y(2\langle w, x \rangle - 1) \leq 0$
- 0-1 loss $l(w, (x, y)) = I[y2\langle w, x \rangle - 1) \leq 0]$
- Define surrogate loss $f_t = [1 - y_t2\langle w, x_t \rangle - 1]_+$
- $f_t$ is convex and for all $w$, $f_t(w) \geq 0$-1 loss
Run Exponentiated Gradient Descent on the sequence of functions $f_t(w)$ with

$$z_t = \begin{cases} 
2y_t x_t, & t \in M \\
0, & \text{otherwise}
\end{cases}$$

to get updates

$$\forall i, w_{t+1} = \begin{cases} 
w_t(i), & y_t 2(w_t, x_t) - 1 \geq 0 \\
w_t(i) e^{-\eta 2y_t x_t(i)}, & \text{otherwise}
\end{cases}$$
Algorithm: Winnow

initialize: \( w_1 = (1/d, ..., 1/d) \)

for \( t = 1, 2, ..., T \)

receive \( x_t \)

predict \( p_t = \text{sign}(2\langle w_t, x_t \rangle - 1) \)

if \( y_t(2\langle w_t, x_t \rangle - 1) \leq 0 \)

\( \forall i, w_{t+1}(i) = w_t(i)e^{-\eta y_t x_t(i)} \)

else \( w_{t+1} = w_t \)
Theorem

Suppose that \textit{Winnow} runs on a sequence \((x_1, y_1, \ldots, x_T, y_T)\), where \(x_t \in \{0, 1\}^d\) for all \(t\). Let \(M,\) be the rounds on which \textit{Winnow} errs and let \(f_t(w) = I_{[i \in M]}[1 - y_t 2\langle w, x_t \rangle - 1]_+\). Then for any \(u \in \{0, 1\}^d\), such that \(\|u\|_1 = k\) it holds that

\[
M \leq \frac{1}{1 - 2\eta} \left( \sum_t f_t(u) + \frac{k \log(d)}{\eta} \right). \tag{16}
\]
Summary

- Derived bounds for FTL–FoReL
- Introduced strongly-convex regularization
- Used different regularizers to derive OGD–EGD using FoReL
- By convexifying 0-1 loss we saw that OGD $\rightarrow$ Perceptron and EGD $\rightarrow$ Winnow