Online Learning and Online Convex Optimization Shai-Shalev Shwartz

Stefanos Poulis

February 6, 2014

Outline

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 $|f(u) - f(w)| \leq L||u - w||$.

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 $f \rightarrow S : \mathbb{R}$ 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

$$\forall t, w_t = \operatorname*{argmin}_{w \in S} \sum_{i=1}^{t-1} f_i(w) \tag{2}$$

Lemma

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

$$Regret_T(u) = \sum_{t=1}^{T} (f_t(w_t) - f_t(u)) \leq \sum_{t=1}^{T} (f_t(w_t) - f_t(w_{t+1})) \tag{3}$$

Proof.

Sketch: Use induction □

Follow-the-Regularized-Leader (FoReL)

Algorithm: Follow-the-Regularized-Leader

$$\forall t, w_t = \operatorname*{argmin}_{w \in S} \sum_{i=1}^{t-1} f_i(w) + R(w) \tag{4}$$

- $R : S \to \mathbb{R}$ is a regularization term
- The goal of regularization is to stabilize the solution

Follow-the-Regularized-Leader

Example

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

$$w_{t+1} = -\eta \sum_{i=1}^{t} z_i = w_t - \eta z_t \tag{5}$$

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,*

$$Regret_T(u) \leq \frac{1}{2\eta}\|u\|_2^2 + \eta \sum_{t=1}^{T} \|z_t\|_2^2. \tag{6}$$

Proof.

Sketch: Run `FTL` on $f_0, f_1, ..., f_T$, where $f_0 = R$
Use gradient updates $\square$

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

$$Regret_T(u) \leq \frac{1}{2\eta}\|u\|_2^2 + \eta \sum_{t=1}^{T} \|z_t\|_2^2. \tag{7}$$

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 $\|.\|$ iff for all $w \in S$ and $z \in \partial f(w)$ we have that $\|z\|_* \leq L$, where $\|.\|_*$ is the dual norm.

Corollary

Consider previous bound for OGD,

$$Regret_T(u) \leq \frac{1}{2\eta}\|u\|_2^2 + \eta \sum_{t=1}^{T} \|z_t\|_2^2. \tag{8}$$

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

$$Regret_T(u) \leq \frac{1}{2\eta}\|u\|_2^2 + \eta T L^2. \tag{9}$$

Strongly Convex Regularizers

A function is strongly convex if it is **strictly** above its tangent

Definition

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

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

Example

$R(w) = \frac{1}{2}\|w\|_2^2$ 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 $S = \{w \in \mathbb{R}^d : w > 0 \wedge \|w\|_1 \leq B\}$.

Analyzing FoReL with Strongly Convex Regularizers

Theorem

Let $f(1), ..., f(T)$ be a sequence of convex functions such that f_t is L_t-Lipschitz with respect to some norm $\|.\|$. 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 σ-strongly-convex with respect to the same norm. Then for all $u \in S$,*

$$Regret_T(u) \leq R(u) - min_{w \in S} R(w) + \frac{TL^2}{\sigma} \tag{11}$$

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

$$w_{t+1} = w_t - \eta z_t \tag{12}$$

- Running `FoReL` with $R(w) = \sum_{i=1}^{d} w_i log(w_i)$ yields `Exponentiated Gradient Descent`,
  with updates

$$w_{t+1}(i) = w_t(i) e^{\eta z_t(i)} \tag{13}$$

Exponentiated Gradient Descent

Algorithm: Exponentiated Gradient Descent (Un-normalized)

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

Theorem

Let $f(1), ..., f(T)$ be a sequence of convex functions such that f_t is L_t-Lipschitz with respect to some norm $\|.\|$. 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 $S = \{w : \|w\|_1 = B \wedge w > 0\} \subset \mathbb{R}^d$. Then,

$$Regret_T(S) \leq \frac{Blog(d)}{\eta} + \eta BTL^2. \tag{14}$$

Proof.

Sketch: Use strong convexity and Holder's inequality. □

Online Classification

Perceptron

- $y \in \{-1, 1\}$
- A weight vector w makes a mistake on an example $(\mathbf{x}, y)$ whenever $sign(\langle w, x \rangle) \neq y$
- 0-1 loss $l(w, (\mathbf{x}, y)) = I_{[y\langle w,x \rangle \leq 0]}$
- 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

Perceptron

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

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

Perceptron

Algorithm: Perceptron

initialize: $w_1 = 0$
for $t = 1, 2, ..., T$
receive x_t
predict $p_t = sign(\langle w_t, x_t \rangle)$
if $y_t(\langle w_t, x_t \rangle) \leq 0$
$w_{t+1} = w_t + y_t x_t$
else $w_{t+1} = w_t$

Perceptron

Theorem

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

$$\mathbb{M} \leq \sum_t f_t(u) + R\|u\|(\sum_t f_t(u))^{\frac{1}{2}} + R^2\|u\|^2 \tag{15}$$

Theorem

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

$$\mathbb{M} \leq \sum_t f_t(u) + R\|u\|(\sum_t f_t(u))^{\frac{1}{2}} + R^2\|u\|^2 \tag{15}$$

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 $l(w, (\mathbf{x}, y)) = I_{[y2\langle w,x \rangle - 1) \leq 0]}$
- Define surrogate loss $f_t = [1 - y_t 2\langle w, x_t \rangle - 1]_+$
- f_t is convex and for all w, $f_t(w) \geq$ 0-1 loss

Winnow

Run `Exponentiated Gradient Descent` on the sequence of functions $f_t(w)$
with

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

to get updates

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

Winnow

Algorithm: Winnow

initialize: $w_1 = (1/d, ..., 1/d)$
for $t = 1, 2, ..., T$
receive x_t
predict $p_t = sign(2\langle w_t, x_t\rangle - 1)$
if $y_t(2\langle w_t, x_t\rangle - 1) \leq 0$
$\forall i, w_{t+1}(i) = w_t(i)e^{-\eta 2 y_t x_t(i)}$
else $w_{t+1} = w_t$

Winnow

Theorem

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

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

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`