A Generalized Online Mirror Descent with Applications to Regression

F. Orabona, K. Crammer, N.C. Bianchi

Presented by Akhoury Shauryam

December 2, 2023

Table of Contents

Definitions

- A function f is called convex if $\forall u, v : f(v) \geq f(u) + \langle \nabla f(u), v - u \rangle$

Definitions

- A function f is called convex if $\forall u, v : f(v) \geq f(u) + \langle \nabla f(u), v - u \rangle$
- A function is α-strongly convex for a norm $\|\cdot\|$ if $\forall u, v : f(v) \geq f(u) + \langle \nabla f(u), v - u \rangle + \frac{\alpha}{2}\|u - v\|^2$

Definitions

- A function f is called convex if
  $\forall u, v : f(v) \geq f(u) + \langle \nabla f(u), v - u \rangle$
- A function is α-strongly convex for a norm $\|\cdot\|$ if
  $\forall u, v : f(v) \geq f(u) + \langle \nabla f(u), v - u \rangle + \frac{\alpha}{2}\|u - v\|^2$
- A function is β-smooth for a norm $\|\cdot\|$ if
  $\forall u, v : f(v) \leq f(u) + \langle \nabla f(u), v - u \rangle + \frac{\beta}{2}\|u - v\|^2$

Fenchel Conjugate

- For a function f, it's Fenchel conjugate f^* is defined as $f^*(u) = \sup(\langle v, u \rangle - f(v))$

Fenchel Conjugate

- For a function f, it's Fenchel conjugate f^* is defined as $f^*(u) = \sup(\langle v, u \rangle - f(v))$
- A vector x is a subgradient of f at v if $\forall u : f(u) - f(v) \geq \langle u - v, x \rangle$, the set of such x is denoted as $\partial f(v)$

Fenchel Conjugate

- For a function f, it's Fenchel conjugate f^* is defined as $f^*(u) = \sup(\langle v, u\rangle - f(v))$
- A vector x is a subgradient of f at v if $\forall u : f(u) - f(v) \geq \langle u - v, x\rangle$, the set of such x is denoted as $\partial f(v)$
- The Fenchel-Young inequality states that if $x \in \partial f(v)$ then $f(v) + f^*(x) = \langle v, x\rangle$

Fenchel Conjugate Properties

- The Fenchel conjugate f^* of an α-strongly convex function f is everywhere differentiable and $\frac{1}{\alpha}$-strongly smooth. This means that, for all $u, v \in X$,

$$f^*(v) \leq f^*(u) + \langle \nabla f^*(u), v - u \rangle + \frac{1}{2\alpha} \|u - v\|^2$$

Fenchel Conjugate Properties

- The Fenchel conjugate f^* of an α-strongly convex function f is everywhere differentiable and $\frac{1}{\alpha}$-strongly smooth. This means that, for all $u, v \in X$,

$$f^*(v) \leq f^*(u) + \langle \nabla f^*(u), v - u \rangle + \frac{1}{2\alpha} \|u - v\|^2$$

- $\nabla f^*(u) = \arg\max_{v \in S} \langle v, u \rangle - f(v)$

Online Convex Optimization

In the online convex optimization protocol, an algorithm sequentially chooses elements from a convex set $S \in X$, each time incurring a certain loss. At each step $t = 1, 2, \ldots$ the algorithm chooses $w_t \in S$ and then observes a convex loss function $l_t : S \Rightarrow R$. The value $l_t(w_t)$ is the loss of the learner at step t, and the goal is to control the regret.

Online Convex Optimization

In the online convex optimization protocol, an algorithm sequentially chooses elements from a convex set $S \in X$, each time incurring a certain loss. At each step $t = 1, 2, \ldots$ the algorithm chooses $w_t \in S$ and then observes a convex loss function $l_t : S \Rightarrow R$. The value $l_t(w_t)$ is the loss of the learner at step t, and the goal is to control the regret.
Regret here is defined as:

$$R_T(u) = \sum_{t=1}^{T} l_t(w_t) - \sum_{t=1}^{T} l_t(u)$$

Online Convex Optimization

In the online convex optimization protocol, an algorithm sequentially chooses elements from a convex set $S \in X$, each time incurring a certain loss. At each step $t = 1, 2, \ldots$ the algorithm chooses $w_t \in S$ and then observes a convex loss function $l_t : S \Rightarrow R$. The value $l_t(w_t)$ is the loss of the learner at step t, and the goal is to control the regret.
Regret here is defined as:

$$R_T(u) = \sum_{t=1}^{T} l_t(w_t) - \sum_{t=1}^{T} l_t(u)$$

So we try to fine tune our algorithm to select w_t which minimises the Regret over all u.

Online Mirror Descent

The standard Online Mirror Descent looks like:

Online Mirror Descent

The standard Online Mirror Descent looks like:

Algorithm OMD

Input: parameter $\eta > 0$, regularization function $R(x)$.

Let y_1 be such that $\nabla R(y_1) = 0$ and $x_1 = \text{argmin } B_R(x \| y_1)$.

for $t = 1$ to T **do**

  Play x_t.

  Observe the loss function f_t and let $\nabla_t = \nabla f_t(x_t)$.

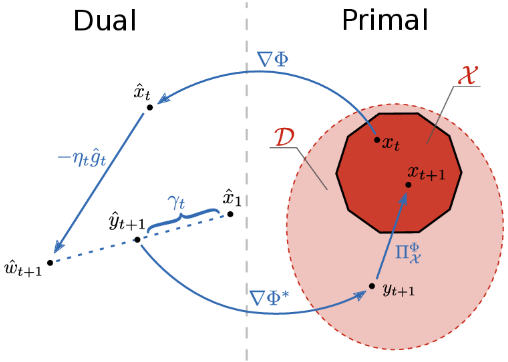
  Update y_{t+1} according to the rule:

Lazy: $\nabla R(y_{t+1}) = \nabla R(y_t) - \eta \nabla_t$

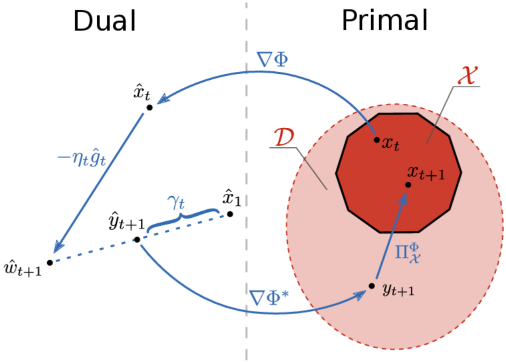
Agile: $\nabla R(y_{t+1}) = \nabla R(x_t) - \eta \nabla_t$

  Project: $x_{t+1} = \text{argmin } B_R(x \| y_{t+1})$

Online Mirror Descent

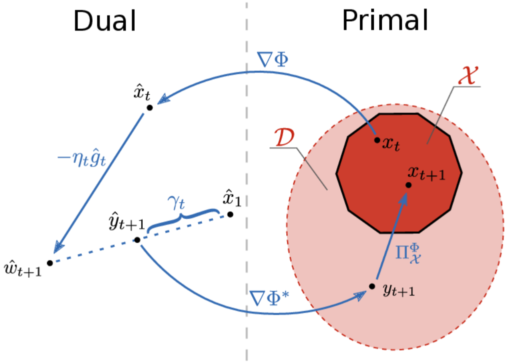


Online Mirror Descent



- In OMD the gradient update is carried out in a "dual" space on the choice of the regularization.

Online Mirror Descent



- In OMD the gradient update is carried out in a "dual" space on the choice of the regularization.
- This transformation enables better bounds depending on what space we begin with.

Generalized Online Mirror Descent

The OMD algorithm is generalized into:

Generalized Online Mirror Descent

The OMD algorithm is generalized into:

Algorithm General Online Mirror Descent

1: **Parameters:** A sequence of strongly convex functions $f_1, f_2, \ldots$ defined on a common convex domain $S \subseteq X$.
2: **Initialize:** $\theta_1 = 0 \in X$
3: **for** $t = 1, 2, \ldots$ **do**
4: Choose $w_t = \nabla f_t^*(\theta_t)$
5: Observe $z_t \in X$
6: Update $\theta_{t+1} = \theta_t + z_t$

Identities for GOMD

Lemma 1

Assume OMD is run with functions $f_1, f_2, \ldots, f_T$ defined on a common convex domain $S \subseteq X$

Identities for GOMD

Lemma 1

Assume OMD is run with functions $f_1, f_2, \ldots, f_T$ defined on a common convex domain $S \subseteq X$
And each f_t is α_t-strongly convex with respect to the norm $\|\cdot\|_t$.
Let $\|\cdot\|_t^*$ be the dual norm of $\|\cdot\|_t$, for $t = 1, 2, \ldots, T$. Then, for any $u \in S$

Identities for GOMD

Lemma 1

Assume OMD is run with functions $f_1, f_2, \ldots, f_T$ defined on a common convex domain $S \subseteq X$
And each f_t is α_t-strongly convex with respect to the norm $\|\cdot\|_t$. Let $\|\cdot\|_t^*$ be the dual norm of $\|\cdot\|_t$, for $t = 1, 2, \ldots, T$. Then, for any $u \in S$

$$\sum_{t=1}^{T} \langle z_t, u - w_t \rangle \leq f_T(u) + \sum_{t=1}^{T} \frac{\|z_t\|_{t,*}^2}{2\alpha_t} + f_t^*(\theta_t) - f_{t-1}^*(\theta_t)$$

Where we set $f_0^*(0) = 0$.

Identities for GOMD

Lemma 1

Assume OMD is run with functions $f_1, f_2, \ldots, f_T$ defined on a common convex domain $S \subseteq X$
And each f_t is α_t-strongly convex with respect to the norm $\|\cdot\|_t$. Let $\|\cdot\|_t^*$ be the dual norm of $\|\cdot\|_t$, for $t = 1, 2, \ldots, T$. Then, for any $u \in S$

$$\sum_{t=1}^{T} \langle z_t, u - w_t \rangle \leq f_T(u) + \sum_{t=1}^{T} \frac{\|z_t\|_{t,*}^2}{2\alpha_t} + f_t^*(\theta_t) - f_{t-1}^*(\theta_t)$$

Where we set $f_0^*(0) = 0$. Moreover, for all $t \geq 1$, we have

Identities for GOMD

Lemma 1

Assume OMD is run with functions $f_1, f_2, \ldots, f_T$ defined on a common convex domain $S \subseteq X$
And each f_t is α_t-strongly convex with respect to the norm $\|\cdot\|_t$. Let $\|\cdot\|_t^*$ be the dual norm of $\|\cdot\|_t$, for $t = 1, 2, \ldots, T$. Then, for any $u \in S$

$$\sum_{t=1}^{T} \langle z_t, u - w_t \rangle \leq f_T(u) + \sum_{t=1}^{T} \frac{\|z_t\|_{t,*}^2}{2\alpha_t} + f_t^*(\theta_t) - f_{t-1}^*(\theta_t)$$

Where we set $f_0^*(0) = 0$. Moreover, for all $t \geq 1$, we have

$$f_t^*(\theta_t) - f_{t-1}^*(\theta_t) \leq f_{t-1}(w_t) - f_t(w_t)$$

Proof for Lemma 1

Let $\Delta_t = f_t^*(\theta_{t+1}) - f_{t-1}^*(\theta_t)$. Then,

Proof for Lemma 1

Let $\Delta_t = f_t^*(\theta_{t+1}) - f_{t-1}^*(\theta_t)$. Then,

$$\sum_{t=1}^{T} \Delta_t = f_T^*(\theta_{T+1}) - f_0^*(\theta_1) = f_T^*(\theta_{T+1})$$

Proof for Lemma 1

Let $\Delta_t = f_t^*(\theta_{t+1}) - f_{t-1}^*(\theta_t)$. Then,

$$\sum_{t=1}^{T} \Delta_t = f_T^*(\theta_{T+1}) - f_0^*(\theta_1) = f_T^*(\theta_{T+1})$$

Since the functions f_t^* are $\frac{1}{\alpha_t}$-strongly smooth with respect to $\|\cdot\|_t^*$, and recalling that $\theta_{t+1} = \theta_t + z_t$,

Proof for Lemma 1

Let $\Delta_t = f_t^*(\theta_{t+1}) - f_{t-1}^*(\theta_t)$. Then,

$$\sum_{t=1}^{T} \Delta_t = f_T^*(\theta_{T+1}) - f_0^*(\theta_1) = f_T^*(\theta_{T+1})$$

Since the functions f_t^* are $\frac{1}{\alpha_t}$-strongly smooth with respect to $\|\cdot\|_t^*$, and recalling that $\theta_{t+1} = \theta_t + z_t$,

$$\Delta_t = f_t^*(\theta_{t+1}) - f_t^*(\theta_t) + f_t^*(\theta_t) - f_{t-1}^*(\theta_t)$$

Proof for Lemma 1

Let $\Delta_t = f_t^*(\theta_{t+1}) - f_{t-1}^*(\theta_t)$. Then,

$$\sum_{t=1}^{T} \Delta_t = f_T^*(\theta_{T+1}) - f_0^*(\theta_1) = f_T^*(\theta_{T+1})$$

Since the functions f_t^* are $\frac{1}{\alpha_t}$-strongly smooth with respect to $\|\cdot\|_t^*$, and recalling that $\theta_{t+1} = \theta_t + z_t$,

$$\Delta_t = f_t^*(\theta_{t+1}) - f_t^*(\theta_t) + f_t^*(\theta_t) - f_{t-1}^*(\theta_t)$$

$$\leq f_t^*(\theta_t) - f_{t-1}^*(\theta_t) + \langle \nabla f_t^*(\theta_t), z_t \rangle + \frac{1}{2\alpha_t} \|z_t\|_{t,*}^2$$

Proof for Lemma 1

Let $\Delta_t = f_t^*(\theta_{t+1}) - f_{t-1}^*(\theta_t)$. Then,

$$\sum_{t=1}^{T} \Delta_t = f_T^*(\theta_{T+1}) - f_0^*(\theta_1) = f_T^*(\theta_{T+1})$$

Since the functions f_t^* are $\frac{1}{\alpha_t}$-strongly smooth with respect to $\|\cdot\|_t^*$, and recalling that $\theta_{t+1} = \theta_t + z_t$,

$$\Delta_t = f_t^*(\theta_{t+1}) - f_t^*(\theta_t) + f_t^*(\theta_t) - f_{t-1}^*(\theta_t)$$

$$\leq f_t^*(\theta_t) - f_{t-1}^*(\theta_t) + \langle \nabla f_t^*(\theta_t), z_t \rangle + \frac{1}{2\alpha_t} \|z_t\|_{t,*}^2$$

$$= f_t^*(\theta_t) - f_{t-1}^*(\theta_t) + \langle w_t, z_t \rangle + \frac{1}{2\alpha_t} \|z_t\|_{t,*}^2.$$

Proof for Lemma 1

The Fenchel-Young inequality implies

$$\sum_{t=1}^{T} \Delta_t = f_T^*(\theta_{T+1}) \geq \langle u, \theta_{T+1} \rangle - f_T(u) = \sum_{t=1}^{T} \langle u, z_t \rangle - f_T(u).$$

Proof for Lemma 1

The Fenchel-Young inequality implies

$$\sum_{t=1}^{T} \Delta_t = f_T^*(\theta_{T+1}) \geq \langle u, \theta_{T+1} \rangle - f_T(u) = \sum_{t=1}^{T} \langle u, z_t \rangle - f_T(u).$$

Summing the last 2 inequalities, we get:

Proof for Lemma 1

The Fenchel-Young inequality implies

$$\sum_{t=1}^{T} \Delta_t = f_T^*(\theta_{T+1}) \geq \langle u, \theta_{T+1} \rangle - f_T(u) = \sum_{t=1}^{T} \langle u, z_t \rangle - f_T(u).$$

Summing the last 2 inequalities, we get:

$$\sum_{t=1}^{T} \langle u, z_t \rangle - f_T(u) \leq \sum_{t=1}^{T} \Delta_t$$

Proof for Lemma 1

So combining the two sum inequalities, we get:

Proof for Lemma 1

So combining the two sum inequalities, we get:

$$\sum_{t=1}^{T}\langle u, z_t\rangle - f_T(u) \leq \sum_{t=1}^{T}\left(f_t^*(\theta_t) - f_{t-1}^*(\theta_t) + \langle w_t, z_t\rangle + \frac{1}{2\beta_t}\|z_t\|_t^2\right).$$

Proof for Lemma 1

So combining the two sum inequalities, we get:

$$\sum_{t=1}^{T} \langle u, z_t \rangle - f_T(u) \leq \sum_{t=1}^{T} \left(f_t^*(\theta_t) - f_{t-1}^*(\theta_t) + \langle w_t, z_t \rangle + \frac{1}{2\beta_t} \|z_t\|_t^2 \right).$$

Rearranging this we get the proof for Lemma 1

Proof for Lemma 1

We now prove the second statement.

Proof for Lemma 1

We now prove the second statement.
Recalling again the definition of w_t, we have that
$f_t^*(\theta_t) = \langle w_t, \theta_t \rangle - f_t(w_t)$.

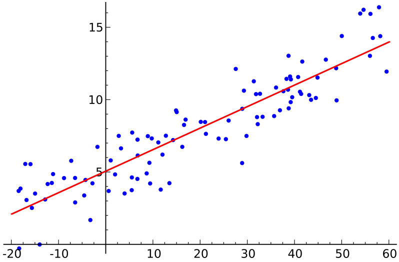
Proof for Lemma 1

We now prove the second statement.
Recalling again the definition of w_t, we have that
$f_t^*(\theta_t) = \langle w_t, \theta_t \rangle - f_t(w_t)$.
On the other hand, the Fenchel-Young inequality implies that
$-f_{t-1}^*(\theta_t) \leq f_{t-1}(w_t) - \langle w_t, \theta_t \rangle$.

Proof for Lemma 1

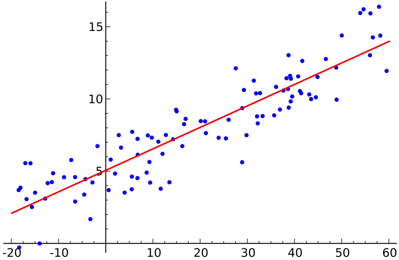
We now prove the second statement.
Recalling again the definition of w_t, we have that
$f_t^*(\theta_t) = \langle w_t, \theta_t \rangle - f_t(w_t)$.
On the other hand, the Fenchel-Young inequality implies that
$-f_{t-1}^*(\theta_t) \leq f_{t-1}(w_t) - \langle w_t, \theta_t \rangle$.
Combining the two, we get
$f_t^*(\theta_t) - f_{t-1}^*(\theta_t) \leq f_{t-1}(w_t) - f_t(w_t)$, as desired.

Linear Regression



- Linear regression is a method that utilizes a linear framework to model the predictive association between a single response variable and one or more explanatory variables.

Linear Regression



- Linear regression is a method that utilizes a linear framework to model the predictive association between a single response variable and one or more explanatory variables.
- We are given pairs of x_t, y_t where $y_t = u^\top x_t + \nu_t$ where ν_t is a random noise and our goal is to recover u.

Online Regression

At time step $t = 1, 2, .., T$, we recieve (x_t, y_t).

Online Regression

At time step $t = 1, 2, .., T$, we recieve (x_t, y_t).
Use w_t based on the history.

Online Regression

At time step $t = 1, 2, .., T$, we recieve (x_t, y_t).
Use w_t based on the history.
Incur a loss based on the Square Loss:

Online Regression

At time step $t = 1, 2, .., T$, we recieve (x_t, y_t).
Use w_t based on the history.
Incur a loss based on the Square Loss:

$$l_t(u) = \frac{(y_t - u^\top x_t)^2}{2}$$

Online Regression

At time step $t = 1, 2, .., T$, we recieve (x_t, y_t).
Use w_t based on the history.
Incur a loss based on the Square Loss:

$$l_t(u) = \frac{(y_t - u^\top x_t)^2}{2}$$

Update w_{t+1} accordingly

Online Regression

At time step $t = 1, 2, .., T$, we recieve (x_t, y_t).
Use w_t based on the history.
Incur a loss based on the Square Loss:

$$l_t(u) = \frac{(y_t - u^\top x_t)^2}{2}$$

Update w_{t+1} accordingly
Here $x_t \in \mathbb{R}^d$ and $y_t \in \mathbb{R}$

Vovk-Azoury-Warmuth Algorithm for Online Regression

For a regulizer constant a, at each time step t, the VAW Algorithm tells us to pick $w_t =$

Vovk-Azoury-Warmuth Algorithm for Online Regression

For a regulizer constant a, at each time step t, the VAW Algorithm tells us to pick $w_t =$

$$= \arg\min_{w} \left(\frac{a}{2}\|w\|^2 + \frac{1}{2}\sum_{s=1}^{t-1}\left(y_s - w^\top x_s\right)^2 + \frac{1}{2}(w^\top x_t)^2 \right)$$

Vovk-Azoury-Warmuth Algorithm for Online Regression

For a regulizer constant a, at each time step t, the VAW Algorithm tells us to pick $w_t =$

$$= \arg\min_w \left(\frac{a}{2}\|w\|^2 + \frac{1}{2}\sum_{s=1}^{t-1}\left(y_s - w^\top x_s\right)^2 + \frac{1}{2}(w^\top x_t)^2 \right)$$

$$= \arg\min_w \left(\tfrac{a}{2}\|w\|^2 + \tfrac{1}{2}\textstyle\sum_{s=1}^{t-1}(w^\top x_s)^2 - \sum_{s=1}^{t-1} y_{s'} w^\top x_s \right)$$

Vovk-Azoury-Warmuth Algorithm for Online Regression

For a regulizer constant a, at each time step t, the VAW Algorithm tells us to pick $w_t =$

$$= \arg\min_w \left(\frac{a}{2}\|w\|^2 + \frac{1}{2}\sum_{s=1}^{t-1}\left(y_s - w^\top x_s\right)^2 + \frac{1}{2}(w^\top x_t)^2 \right)$$

$$= \arg\min_w \left(\tfrac{a}{2}\|w\|^2 + \tfrac{1}{2}\textstyle\sum_{s=1}^{t-1}(w^\top x_s)^2 - \sum_{s=1}^{t-1} y_{s'} w^\top x_s \right)$$
$$= \arg\min_w \left(\tfrac{1}{2} w^\top \left(aI + \textstyle\sum_{i=1}^{t-1} x_s x_s^\top \right) w - \sum_{s=1}^{t-1} y_s w^\top x_s \right)$$

Vovk-Azoury-Warmuth Algorithm for Online Regression

For a regulizer constant a, at each time step t, the VAW Algorithm tells us to pick $w_t =$

$$= \arg\min_{w} \left(\frac{a}{2}\|w\|^2 + \frac{1}{2}\sum_{s=1}^{t-1}\left(y_s - w^\top x_s\right)^2 + \frac{1}{2}(w^\top x_t)^2 \right)$$

$$= \arg\min_w \left(\tfrac{a}{2}\|w\|^2 + \tfrac{1}{2}\textstyle\sum_{s=1}^{t-1}(w^\top x_s)^2 - \sum_{s=1}^{t-1} y_{s'} w^\top x_s \right)$$
$$= \arg\min_w \left(\tfrac{1}{2} w^\top \left(aI + \textstyle\sum_{i=1}^{t-1} x_s x_s^\top \right) w - \sum_{s=1}^{t-1} y_s w^\top x_s \right)$$
$$= \left(aI + \textstyle\sum_{s=1}^{t-1} x_s x_s^\top \right)^{-1} \sum_{i=1}^{t-1} y_i \cdot x_i$$

Fitting VAW to GOMD

- Now, by letting $A_0 = aI$, $A_t = A_{t-1} + x_t x_t^\top$ for $t \geq 1$, and $z_s = y_s \cdot x_s$, we obtain the OMD update $w_t = A_t^{-1}\theta_t = \nabla f_t^*(\theta_t)$

Fitting VAW to GOMD

- Now, by letting $A_0 = aI$, $A_t = A_{t-1} + x_t x_t^\top$ for $t \geq 1$, and $z_s = y_s \cdot x_s$, we obtain the OMD update $w_t = A_t^{-1}\theta_t = \nabla f_t^*(\theta_t)$
- Where $f_t(u) = \frac{1}{2}u^\top A_t u$ and $f_t^*(\theta) = \frac{1}{2}\theta^\top A_t^{-1}\theta$.

Fitting VAW to GOMD

- Now, by letting $A_0 = aI$, $A_t = A_{t-1} + x_t x_t^\top$ for $t \geq 1$, and $z_s = y_s \cdot x_s$, we obtain the OMD update $w_t = A_t^{-1}\theta_t = \nabla f_t^*(\theta_t)$
- Where $f_t(u) = \frac{1}{2}u^\top A_t u$ and $f_t^*(\theta) = \frac{1}{2}\theta^\top A_t^{-1}\theta$.
- The regret bound of this algorithm is recovered from Lemma 1 by noting that f_t is 1-strongly convex with respect to the norm $\|u\|_t = \sqrt{u^\top A_t u}$.

Fitting VAW to GOMD

- Now, by letting $A_0 = aI$, $A_t = A_{t-1} + x_t x_t^\top$ for $t \geq 1$, and $z_s = y_s \cdot x_s$, we obtain the OMD update $w_t = A_t^{-1}\theta_t = \nabla f_t^*(\theta_t)$
- Where $f_t(u) = \frac{1}{2}u^\top A_t u$ and $f_t^*(\theta) = \frac{1}{2}\theta^\top A_t^{-1}\theta$.
- The regret bound of this algorithm is recovered from Lemma 1 by noting that f_t is 1-strongly convex with respect to the norm $\|u\|_t = \sqrt{u^\top A_t u}$.

Hence, the regret $R_T(u)$ is controlled as follows:

Regret for GOMD-VAW

$$R_T(u) = \frac{1}{2}\sum_{t=1}^{T}(y_t - w_t^\top x_t)^2 - \frac{1}{2}\sum_{t=1}^{T}(y_t - u^\top x_t)^2$$

Regret for GOMD-VAW

$$R_T(u) = \frac{1}{2}\sum_{t=1}^{T}(y_t - w_t^\top x_t)^2 - \frac{1}{2}\sum_{t=1}^{T}(y_t - u^\top x_t)^2$$

$$= \sum_{t=1}^{T}(y_t u^\top x_t - y_t w_t^\top x_t) - f_T(u) + \frac{a}{2}\|u\|_2^2 + \frac{1}{2}\sum_{t=1}^{T}(w_t^\top x_t)^2$$

Regret for GOMD-VAW

$$R_T(u) = \frac{1}{2}\sum_{t=1}^{T}(y_t - w_t^\top x_t)^2 - \frac{1}{2}\sum_{t=1}^{T}(y_t - u^\top x_t)^2$$

$$= \sum_{t=1}^{T}(y_t u^\top x_t - y_t w_t^\top x_t) - f_T(u) + \frac{a}{2}\|u\|_2^2 + \frac{1}{2}\sum_{t=1}^{T}(w_t^\top x_t)^2$$

$$\leq f_T(u)+\sum_{t=1}^{T} y_t^2\|x_t\|_{2t,*}^2+f_t^*(\theta_t)-f_{t-1}^*(\theta_t)-f_T(u)+\frac{a}{2}\|u\|_2^2+\frac{1}{2}\sum_{t=1}^{T}(w_t^\top x_t)^2$$

Regret for GOMD-VAW

$$R_T(u) = \frac{1}{2}\sum_{t=1}^{T}(y_t - w_t^\top x_t)^2 - \frac{1}{2}\sum_{t=1}^{T}(y_t - u^\top x_t)^2$$

$$= \sum_{t=1}^{T}(y_t u^\top x_t - y_t w_t^\top x_t) - f_T(u) + \frac{a}{2}\|u\|_2^2 + \frac{1}{2}\sum_{t=1}^{T}(w_t^\top x_t)^2$$

$$\leq f_T(u) + \sum_{t=1}^{T} y_t^2 \|x_t\|_{2t,*}^2 + f_t^*(\theta_t) - f_{t-1}^*(\theta_t) - f_T(u) + \frac{a}{2}\|u\|_2^2 + \frac{1}{2}\sum_{t=1}^{T}(w_t^\top x_t)^2$$

$$\leq \frac{a}{2}\|u\|^2 + \frac{Y^2}{2}\sum_{t=1}^{T} x_t^\top A_t^{-1} x_t, \text{ where, } \{Y = max\|y_t\|_t\}$$

Regret for GOMD-VAW

$$R_T(u) = \frac{1}{2}\sum_{t=1}^{T}(y_t - w_t^\top x_t)^2 - \frac{1}{2}\sum_{t=1}^{T}(y_t - u^\top x_t)^2$$

$$= \sum_{t=1}^{T}(y_t u^\top x_t - y_t w_t^\top x_t) - f_T(u) + \frac{a}{2}\|u\|_2^2 + \frac{1}{2}\sum_{t=1}^{T}(w_t^\top x_t)^2$$

$$\leq f_T(u) + \sum_{t=1}^{T} y_t^2\|x_t\|_{2t,*}^2 + f_t^*(\theta_t) - f_{t-1}^*(\theta_t) - f_T(u) + \frac{a}{2}\|u\|_2^2 + \frac{1}{2}\sum_{t=1}^{T}(w_t^\top x_t)^2$$

$$\leq \frac{a}{2}\|u\|^2 + \frac{Y^2}{2}\sum_{t=1}^{T} x_t^\top A_t^{-1} x_t, \text{ where, } \{Y = max\|y_t\|_t\}$$

$$\text{since } f_t^*(\theta_t) - f_{t-1}^*(\theta_t) \leq f_{t-1}(w_t) - f_t(w_t) = -\frac{1}{2}(w_t^\top x_t)^2$$

Adaptive Filtering Regret

We use a new variant of Regret named the AF-Regret, defined as follows:

Adaptive Filtering Regret

We use a new variant of Regret named the AF-Regret, defined as follows:

$$R_T^{AF}(u) = \sum_{t=1}^{T} (w_t^\top x_t - u^\top x_t)^2$$

Notice that $R_T(u) + \frac{1}{2} R_T^{AF}(u) =$

Adaptive Filtering Regret

We use a new variant of Regret named the AF-Regret, defined as follows:

$$R_T^{AF}(u) = \sum_{t=1}^{T} (w_t^\top x_t - u^\top x_t)^2$$

Notice that $R_T(u) + \frac{1}{2} R_T^{AF}(u) =$

$$\sum_{t=1}^{T} \left((y_t - w_t^\top x_t)^2 - (y_t - u^\top x_t)^2 + \frac{1}{2} (w_t^\top x_t - u^\top x_t)^2 \right)$$

Adaptive Filtering Regret

We use a new variant of Regret named the AF-Regret, defined as follows:

$$R_T^{AF}(u) = \sum_{t=1}^{T} (w_t^\top x_t - u^\top x_t)^2$$

Notice that $R_T(u) + \frac{1}{2} R_T^{AF}(u) =$

$$\sum_{t=1}^{T} \left((y_t - w_t^\top x_t)^2 - (y_t - u^\top x_t)^2 + \frac{1}{2}(w_t^\top x_t - u^\top x_t)^2 \right)$$

$$= \sum_{t=1}^{T} \left((y_t - w_t^\top x_t) u^\top x_t - (y_t - w_t^\top x_t) w_t^\top x_t \right)$$

Adaptive Filtering Regret

We use a new variant of Regret named the AF-Regret, defined as follows:

$$R_T^{AF}(u) = \sum_{t=1}^{T} (w_t^\top x_t - u^\top x_t)^2$$

Notice that $R_T(u) + \frac{1}{2} R_T^{AF}(u) =$

$$\sum_{t=1}^{T} \left((y_t - w_t^\top x_t)^2 - (y_t - u^\top x_t)^2 + \frac{1}{2} (w_t^\top x_t - u^\top x_t)^2 \right)$$

$$= \sum_{t=1}^{T} \left((y_t - w_t^\top x_t) u^\top x_t - (y_t - w_t^\top x_t) w_t^\top x_t \right)$$

$$= \sum_{t=1}^{T} (u - w_t)^\top z_t$$

AF Regret Bounding

We set z_t to equal $-(y_t - w_t^\top x_t)x_t$

AF Regret Bounding

We set z_t to equal $-(y_t - w_t^\top x_t)x_t$
Now, pick any function f which is 1-strongly convex with respect to some norm $\|\cdot\|$, and let $f_t(u) = X_t^2 f(u)$, where $X_t = \max_{s \le t} \|x_s\|_*$

AF Regret Bounding

We set z_t to equal $-(y_t - w_t^\top x_t)x_t$
Now, pick any function f which is 1-strongly convex with respect to some norm $\|\cdot\|$, and let $f_t(u) = X_t^2 f(u)$, where $X_t = \max_{s \leq t} \|x_s\|_*$
Lemma 1 then immediately implies that

$$\sum_{t=1}^{T} \langle u - w_t, z_t \rangle \leq f_T(u) + \frac{1}{2} \sum_{t=1}^{T} (y_t - w_t^\top x_t)^2$$

AF Regret Bounding

We set z_t to equal $-(y_t - w_t^\top x_t)x_t$
Now, pick any function f which is 1-strongly convex with respect to some norm $\|\cdot\|$, and let $f_t(u) = X_t^2 f(u)$, where $X_t = \max_{s \leq t} \|x_s\|_*$
Lemma 1 then immediately implies that

$$\sum_{t=1}^{T} \langle u - w_t, z_t \rangle \leq f_T(u) + \frac{1}{2} \sum_{t=1}^{T} (y_t - w_t^\top x_t)^2$$

Where we used the X_t^2-strong convexity of f_t and the fact that $f_t \geq f_{t-1}$.

AF Regret Bounding

Combining previous inequalities and the bound for $R_T(u)$, we get:

$$R_T^{AF}(u) \leq 2X_T^2 f(u) + \sum_{t=1}^{T} (y_t - u^\top x_t)^2$$

Conclusions

- The Authors have generalized Online Mirror Descent for time-varying regularizer.

Conclusions

- The Authors have generalized Online Mirror Descent for time-varying regularizer.
- They modeled Linear Regression into an Online Learning scenario to analyze regret bounds.

Conclusions

- The Authors have generalized Online Mirror Descent for time-varying regularizer.
- They modeled Linear Regression into an Online Learning scenario to analyze regret bounds.
- Relating to the VAW Algorithm, they fit the GOMD model and found tighter bounds for the Regret and AF-Regret.

Conclusions

- The Authors have generalized Online Mirror Descent for time-varying regularizer.
- They modeled Linear Regression into an Online Learning scenario to analyze regret bounds.
- Relating to the VAW Algorithm, they fit the GOMD model and found tighter bounds for the Regret and AF-Regret.
- They further expanded upon this to work for Classification models using GOMD.

Thank You!