A generalized online mirror descent with applications to classification and regression

Francesco Orabona et al. Machine Learning, 2015.

presented by Boyoung Kim

May 31, 2019

Online convex optimization

- $\mathbb{X}$ be any finite-dimensional linear space equipped with inner product $\langle \cdot, \cdot \rangle$. eg) $\mathbb{X} = \mathbb{R}^d$ where $\langle \cdot, \cdot \rangle$ is the vector dot product
- At each step $t = 1, 2, \ldots$ the algorithm chooses $w_t \in S \subseteq \mathbb{X}$ and then observes a convex loss function $\ell_t : S \to \mathbb{R}$, the goal is to control the regret

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

for all $u \in S$.

- In these settings for a fixed but unknown example $(x_t, y_t) \in \mathbb{X} \times \mathbb{R}$ the loss suffered at step t is defined as $\ell_t(\boldsymbol{w}_t) = \ell(\langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle, y_t)$.

Further notation and definitions

- We consider functions f that are closed and convex with domain $S \subseteq \mathbb{X}$.
- Its Fenchel conjugate $f^* : \mathbb{X} \to \mathbb{R}$ is defined by

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

  - The domain of f^* is always $\mathbb{X}$.
  - $f^{**} = f$
- $\|u\|$: A generic norm of a vector $u \in \mathbb{X}$.
  - dual $\|\cdot\|_*$ is the norm defined by $\|\boldsymbol{v}\|_* = \sup_u \{\langle \boldsymbol{u}, \boldsymbol{v} \rangle : \|\boldsymbol{u}\| \leq 1\}$.
  - The Fenchel-Young inequality states that $f(\boldsymbol{u}) + f^*(\boldsymbol{v}) \geq \langle \boldsymbol{u}, \boldsymbol{v} \rangle$ for all $\boldsymbol{v}, \boldsymbol{u}$
- A vector x is a subgradient of a convex function f at v if $f(u) - f(v) \geq \langle u - v, x \rangle$ for any u in the domain of f.
  - $\partial f(v)$: the set of all the subgradients of f at v
  - $\nabla f(v)$: the gradient of f at v
  - For all $x \in \partial f(v)$ we have that $f(v) + f^*(x) = \langle v, x \rangle$

Further notation and definitions

- A function f is β -strongly convex with respect to a norm $\|\cdot\|$ if for any u, v in its domain, and any $x \in \partial f(u)$

$$f(v) \geq f(u) + \langle x, v-u \rangle + \frac{\beta}{2}\|u-v\|^2$$

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

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

- A further property of strongly convex functions $f : S \to \mathbb{R}$ is the following:
    - For all $u \in \mathbb{X}$,

$$\nabla f^*(u) = \operatorname*{argmax}_{v \in S}(\langle v, u \rangle - f(v)) \tag{2}$$

    - This implies

$$f(\nabla f^*(\boldsymbol{u})) + f^*(\boldsymbol{u}) = \langle \nabla f^*(\boldsymbol{u}), \boldsymbol{u} \rangle \tag{3}$$

Online mirror descent

- The standard OMD algorithm sets
  - $\boldsymbol{w}_t = \nabla f^* (\boldsymbol{\theta}_t)$ where f is a strongly convex regularizer
  - θ_t is updated using subgradient descent: $\theta_{t+1} = \theta_t - \eta \ell_t'$ for $\eta > 0$ and $\ell_t' \in \partial \ell_t (w_t)$

- We genaralize OMD in two ways :
  - We allow f to change over time
  - We do not necessarily use the subgradient of the loss to update θ_t

---

Algorithm 1 Online Mirror Descent

1: **Parameters:** A sequence of strongly convex functions $f_1, f_2, \ldots$ defined on a common convex domain $S \subseteq \mathbb{X}$.
2: **Initialize:** $\boldsymbol{\theta}_1 = \mathbf{0} \in \mathbb{X}$
3: **for** $t = 1, 2, \ldots$ **do**
4:    Choose $\boldsymbol{w}_t = \nabla f_t^*(\boldsymbol{\theta}_t)$
5:    Observe $z_t \in \mathbb{X}$
6:    Update $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + z_t$
7: **end for**

Online mirror descent

Lemma 1

Assume OMD is run with functions $f_1, f_2, \ldots, f_T$ defined on a common convex domain $S \subseteq \mathbb{X}$ and such that each f_t is β_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 $u \in S$,

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

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

$$f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) \leq f_{t-1}(w_t) - f_t(\boldsymbol{w}_t) \tag{4}$$

Online mirror descent

Proof.

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 f_t^* are $\frac{1}{\beta_t}$ -strongly smooth with respect to $\|\cdot\|_{t,*}$, and $\theta_{t+1} = \theta_t + z_t$,

$$\begin{aligned}\Delta_t &= f_t^*(\boldsymbol{\theta}_{t+1}) - f_t^*(\boldsymbol{\theta}_t) + f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) \\ &\leq f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) + \langle \nabla f_t^*(\boldsymbol{\theta}_t), z_t\rangle + \frac{1}{2\beta_t}\|z_t\|_{t,*}^2 \\ &= f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) + \langle \boldsymbol{w}_t, z_t\rangle + \frac{1}{2\beta_t}\|z_t\|_{t,*}^2\end{aligned}$$

The Fenchel-Young inequality implies

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

Combining the upper and lower bound on Δ_t and summing over t we get the first statement. $\square$

Online mirror descent

Proof.

(continued) We now prove the second statement. Recalling again $w_t = \nabla f_t^* (\boldsymbol{\theta}_t)$, we have that (3) implies

$$f_t^* (\theta_t) = \langle w_t, \theta_t \rangle - f_t (w_t) .$$

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

Online mirror descent

Regret bounds for OMD applied to three different classes of time-varying regularizers. While the composite setting $(\ell_t(\cdot) + F(\cdot))$ is considered more difficult than the standard one, here we show that this setting can be efficiently solved using OMD with a specific choice of the timevarying regularizer.

Corollary 1

Let S a convex set, $F : S \to \mathbb{R}$ be a convex function, and let $g_1, g_2, \ldots, g_T$ be a sequence of convex functions $g_t : S \to \mathbb{R}$ such that $g_t(u) \leq g_{t+1}(u)$ for all $t = 1, 2, \ldots, T$ and all $u \in S$. Fix $\eta > 0$ and assume $f_t = g_t + \eta t F$ are β_t -strongly convex w.r.t. $\|\cdot\|_t$. For each $t = 1, 2, \ldots, T$ let $\|\cdot\|_{t,*}$ be the dual norm of $\|\cdot\|_t$ is run on the input sequence $z_t = -\eta\ell_t'$ for some $\ell_t' \in \partial\ell_t(w_t)$, then

$$\sum_{t=1}^{T}\left(\ell_t\left(w_t\right) + F\left(w_t\right)\right) - \sum_{t=1}^{T}\left(\ell_t(u) + F(u)\right) \leq \frac{g_T(u)}{\eta} + \eta\sum_{t=1}^{T}\frac{\|\ell_t'\|_{t,*}^2}{2\beta_t} \qquad (5)$$

for all $u \in S$.

Corollary 1 (continued)

Moreover, if $f_t = g\sqrt{t} + \eta t F$ *where* $g : S \to \mathbb{R}$ *is* β *-strongly convex w.r.t.* $\|\cdot\|$ *, then*

$$\sum_{t=1}^{T} \left(\ell_t\left(w_t\right) + F\left(w_t\right)\right) - \sum_{t=1}^{T} \left(\ell_t(u) + F(u)\right) \leq \sqrt{T} \left(\frac{g(u)}{\eta} + \frac{\eta}{\beta} \max_{t \leq T} \left\| \ell_t' \right\|_*^2 \right) \quad (6)$$

for all $u \in S$.

Finally, if $f_t = tF$, *where* F *is* β *-strongly convex w.r.t.* $\|\cdot\|$ *, then*

$$\sum_{t=1}^{T} \left(\ell_t\left(w_t\right) + F\left(w_t\right)\right) - \sum_{t=1}^{T} \left(\ell_t(u) + F(u)\right) \leq \max_{t \leq T} \left\| \ell_t' \right\|_*^2 \frac{(1 + \ln T)}{2\beta} \quad (7)$$

for all $u \in S$.

Online mirror descent

Proof.

By convexity, $\ell_t\left(w_t\right)-\ell_t(u) \leq \frac{1}{\eta}\left\langle z_t, u-w_t\right\rangle$. Using Lemma 1 we have,

$$\sum_{t=1}^{T}\left\langle z_t, u-w_t\right\rangle \leq g_T(u)+\eta T F(u)+\eta^2 \sum_{t=1}^{T} \frac{\left\|\ell_t^{\prime}\right\|_{t, *}^2}{2 \beta_t}+\eta \sum_{t=1}^{T}\left((t-1) F\left(w_t\right)-t F\left(w_t\right)\right)$$

where we used the fact that the terms $g_{t-1}\left(w_t\right)-g_t\left(w_t\right)$ are nonpositive as per our assumption. Reordering terms we obtain (5).

In order to obtain (6) it is sufficient to note that f_t is $\beta\sqrt{t}$ -strongly convex and the inequality $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ concludes the proof.

Finally, bound (7) is proven by observing that $f_t = tF$ is βt -strongly convex and the inequality $\sum_{t=1}^{T} \frac{1}{t} \leq 1+\ln T$ concludes the proof. □

Online regression with square loss

- We apply Lemma 1 to recover known regret bounds for online regression with the square loss.
- For simplicity, we set $\mathbb{X} = \mathbb{R}^d$ and let the inner product $\langle u, x \rangle = u^\top x$
- We also set $\ell_t(u) = \frac{1}{2}\left(y_t - u^\top x_t\right)^2$ for examples $(x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$

Online regression with square loss

- Specialize OMD to the Vovk–Azoury–Warmuth(VAW) algorithm for online regression
- VAW algorithm predicts with, at each time step t,

$$
\begin{aligned}
w_t &= \underset{w}{\operatorname{argmin}} \frac{a}{2}\|w\|^2 + \frac{1}{2}\sum_{s=1}^{t-1}\left(y_s - w^\top x_s\right)^2 + \frac{1}{2}\left(w^\top x_t\right)^2 \\
&= \underset{w}{\operatorname{argmin}} \frac{1}{2} w^\top \left(aI + \sum_{i=1}^{t} x_s x_s^\top\right) w - \sum_{s=1}^{t-1} y_s w^\top x_s \\
&= \left(aI + \sum_{s=1}^{t} x_s x_s^\top\right)^{-1} \sum_{i=1}^{t-1} y_s x_s
\end{aligned}
$$

- Now, by letting $A_0 = aI_d, A_t = A_{t-1} + x_t x_t^\top$ for $t \geq 1$, and $z_s = y_s x_s$, we obtain the OMD update $w_t = A_t^{-1}\theta_t = \nabla f_t^*\left(\theta_t\right)$, where $f_t(u) = \frac{1}{2}u^\top A_t u$ and $f_t^*(\theta) = \frac{1}{2}\theta^\top A_t^{-1}\theta$.
- Note) z_t is not equal to the negative gradient of the square loss.

Online regression with square loss

- The regret bound of this algorithm is recovered from Lemma 1 by noting that f_t is 1-strongly convex w.r.t. the norm $\|u\|_t = \sqrt{u^\top A_t u}$. $\|u\|_{t,*} = \sqrt{u^\top A_t^{-1} u}$.
- Hence, the regret bound is

$$\begin{aligned}
R_T(u) &= \frac{1}{2}\sum_{t=1}^{T}\left(y_t - w_t^\top x_t\right)^2 - \frac{1}{2}\sum_{t=1}^{T}\left(y_t - u^\top x_t\right)^2 \\
&= \sum_{t=1}^{T}\left(y_t u^\top x_t - y_t w_t^\top x_t\right) - f_T(u) + \frac{a}{2}\|u\|^2 + \frac{1}{2}\sum_{t=1}^{T}\left(w_t^\top x_t\right)^2 \\
&\leq f_T(u) + \sum_{t=1}^{T}\left(\frac{y_t^2\,\|x_t\|_{t,*}^2}{2} + f_t^*\left(\theta_t\right) - f_{t-1}^*\left(\theta_t\right)\right) - f_T(u) + \frac{a}{2}\|u\|^2 \\
&\quad + \frac{1}{2}\sum_{t=1}^{T}\left(w_t^\top x_t\right)^2 \\
&\leq \frac{a}{2}\|u\|^2 + \frac{Y^2}{2}\sum_{t=1}^{T} x_t^\top A_t^{-1} x_t
\end{aligned}$$

since $f_t^*\left(\theta_t\right) - f_{t-1}^*\left(\theta_t\right) \leq f_{t-1}\left(w_t\right) - f_t\left(w_t\right) = -\frac{1}{2}\left(w_t^\top x_t\right)^2$, and where $Y = \max_t |y_t|$.

Scale-invariant algorithms

- We introduce two new scale invariant algorithms for online linear regression with an arbitrary convex and Lipschitz loss function.
- Let $\mathbb{X} = \mathbb{R}^d$ and let the inner product $\langle u, x \rangle$ be the standard dot product $u^\top x$

Scale-invariant algorithms

- We assume
  - For loss $\ell_t(w) = \ell\left(w^\top x_t, y_t\right)$, ℓ is L -Lipschitz for each y_t and convex.
  - OMD is run with $z_t = -\eta \ell'_t$ where, as usual, $\ell'_t \in \partial \ell_t\left(w_t\right)$.
  - In the rest of this section, the following notation is used:

    $b_{t,i} = \max_{s=1,\ldots,t}\left|x_{s,i}\right|$, $m_t = \max_{s=1,\ldots,t}\left\|x_s\right\|_0$, $p_t = 2\ln m_t$, and

$$\beta_t = \sqrt{eL^2\left(p_t - 1\right) + \sum_{s=1}^{t-1}\left(p_s - 1\right)\left(\sum_{i=1}^{d}\left(\frac{\left|\ell'_{s,i}\right|}{b_{s,i}}\right)^{p_s}\right)^{2/p_s}}$$

- The time-varying regularizers we consider are defined as follows,

$$f_t(u) = \frac{\beta_t}{2}\left(\sum_{i=1}^{d}\left(\left|u_i\right| b_{t,i}\right)^{q_t}\right)^{2/q_t} \quad \text{for } q_t = \frac{p_t}{p_t - 1} \tag{8}$$

$$f_t(u) = \frac{\sqrt{d}}{2}\left(\sum_{i=1}^{d}\left(\left|u_i\right| b_{t,i}\right)^2 \sqrt{L^2 + \sum_{s=1}^{t-1}\left(\frac{\ell'_{s,i}}{b_{s,i}}\right)^2}\right) \tag{9}$$

Scale-invariant algorithms

- OMD update :
  - For regularizers of type (8) we have

$$(\nabla f_t^*(\theta))_j = \frac{1}{\beta_t (p_t - 1)} \left(\sum_{i=1}^{d} \left(\frac{|\theta_i|}{b_{t,i}} \right)^{p_t} \right)^{2/p_i - 1} \frac{|\theta_j|^{p_t - 1}}{b_{t,j}^{p_t}} \operatorname{sign}(\theta_j)$$

  - For regularizers of type (9) we have

$$(\nabla f_t^*(\theta))_j = \frac{\theta_j}{b_{t,j}^2 \sqrt{d} \sqrt{L^2 + \sum_{s=1}^{t-1} \left(\frac{\ell'_{s,j}}{b_{s,j}} \right)^2}}$$

- Computation : using the fact that if $g(w) = af(w)$, then $g^*(\theta) = af^*\left(\frac{\theta}{a}\right)$, and Lemma 2 in the appendix.
- Note) $\boldsymbol{w}_t^\top \boldsymbol{x}_t$ is invariant to the rescaling of individual features.

Scale-invariant algorithms

We prove the following regret bounds.

Theorem 1

If OMD is run using regularizers of type (8), then for any $u \in \mathbb{R}^d$

$$R_T(u) \leq L\sqrt{e(T+1)\,(2\ln m_T - 1)}\left(\frac{1}{2\eta}\left(\sum_{i=1}^{d}|u_i|\, b_{T,i}\right)^2 + \eta\right)$$

If OMD is run using regularizers of type (9), then for any $u \in \mathbb{R}^d$

$$R_T(u) \leq L\sqrt{d(T+1)}\left(\frac{1}{2\eta}\sum_{i=1}^{d}\left(u_i b_{T,i}\right)^2 + \eta\right).$$

- Note) both bounds are invariant with respect to arbitrary scaling of individual coordinates of the data points x_t : if the i th feature is rescaled $x_{t,i} \to c x_{t,i}$ for all t, then a corresponding rescaling $u_i \to u_i / c$, leaves the bounds unchanged.

Proof.

For the first algorithm, note that $m_t^{2/p_t} = e$, and setting $q_t = \left(1 - \frac{1}{p_t}\right)^{-1}$, we have $q_t\left(1 - p_t\right) = -p_t$. Further note that $f_t^*\left(\theta_t\right) - f_{t-1}^*\left(\theta_t\right) \leq f_{t-1}\left(w_t\right) - f_t\left(w_t\right) \leq 0$, where $f_{t-1} \leq f_t$ because q_t is decreasing, $b_{t,i}$ is increasing, and β_t is also increasing. Hence, using the convexity of ℓ_t and Lemma 1, we may write

$$
\begin{aligned}
R_T(\boldsymbol{u}) &\leq \sum_{t=1}^{T} \left(\ell_t'\right)^\top \left(\boldsymbol{w}_t - \boldsymbol{u}\right) \\
&\leq \frac{\beta_T}{2\eta} \left(\sum_{i=1}^{d} \left(\left|u_i\right| b_{T,i}\right)^{q_T}\right)^{2/q_T} + \eta \sum_{t=1}^{T} \frac{1}{2\beta_t\left(q_t - 1\right)} \left(\sum_{i=1}^{d} \frac{\left|\ell_{t,i}'\right|^{p_t}}{b_{t,i}^{p_t}}\right)^{2/p_t}
\end{aligned}
$$

For the rest of the proof, reference the paper. □

Binary classification

- We show that mistake bound of special case of online convex optimization
- Let $\mathbb{X}$ be any finite-dimensional inner product space.
- $(x_t, y_t) \in \mathbb{X} \times \{-1, +1\}$, let $\ell_t(w)$ be hinge loss $[1 - y_t \langle w, x_t \rangle]_+$
  - It is easy to verify that the hinge loss satisfies :

    If $\ell_t(w) > 0$ then $\ell_t(u) \geq 1 + \langle u, \ell_t' \rangle$ for all $u, w \in \mathbb{R}^d$ with $\ell_t' \in \partial \ell_t(w)$

    (this is used for the proof of Lemma 3)
  - Note that when $\ell_t(w) > 0$, $\partial \ell_t(w)$ is the singleton $\{\nabla \ell_t(w)\}$

Binary classification

- Set $z_t = -\eta_t \ell'_t$ if $\ell_t(w_t) > 0$, and $z_t = 0$ otherwise.
- Made a prediction mistake is defined by the condition $y_t w_t^\top x_t \leq 0$ or, equivalently, by $\ell_t(w_t) \geq 1$
  - $\mathcal{M}$: the subset of steps t such that $y_t w_t^\top x_t \leq 0$ and by M its cardinality
  - $\mathcal{U}$: the set of margin error steps; that is, steps t where $y_t w_t^\top x_t > 0$ and $\ell_t(w_t) > 0$ and we use U to denote the cardinality of $\mathcal{U}$

Binary classification

Let $L(u) = \sum_{t=1}^{T} [1 - y_t \langle u, x_t \rangle]_+$ for $u \in \mathbb{X}$.

Corollary 2

Assume OMD is run with $f_t = f$, *where* f *has domain* $\mathbb{X}$, *is* β *-strongly convex with respect to the norm* $\|\cdot\|$, *and satisfies* $f(\lambda u) \leq \lambda^2 f(u)$ *for all* $\lambda \in \mathbb{R}$ *and all* $u \in \mathbb{X}$. *Further assume the input sequence is* $z_t = \eta_t y_t x_t$ *for some* $0 \leq \eta_t \leq 1$ *such that* $\eta_t = 1$ *whenever* $y_t \langle w_t, x_t \rangle \leq 0$. *Then, for all* $T \geq 1$

$$M \leq \underset{u \in \mathbb{X}}{\operatorname{argmin}} L(u) + D + \frac{2}{\beta} f(u) X_T^2 + X_T \sqrt{\frac{2}{\beta} f(u) L(u)}$$

where $M = |\mathcal{M}|, X_t = \max_{i=1,\ldots,t} \|x_i\|_*$ *and*

$$D = \sum_{t \in \mathcal{U}} \eta_t \left(\frac{\eta_t \|\boldsymbol{x}_t\|_*^2 + 2\beta y_t \langle w_t, x_t \rangle}{X_t^2} - 2 \right)$$

Proof.

Fix any $u \in \mathbb{X}$. Using the second bound of Lemma 3 in the "Appendix", with the assumption $\eta_t = 1$ when $t \in \mathcal{M}$, we get

$$M \le L(u) + \sqrt{2f(u)}\sqrt{\sum_{t\in\mathcal{M}} \frac{\|x_t\|_*^2}{\beta} + \sum_{t\in\mathcal{U}} \left(\frac{\eta_t^2}{\beta} \|x_t\|_*^2 + 2\eta_t y_t \langle w_t, x_t\rangle \right)} - \sum_{t\in\mathcal{U}} \eta_t$$

$$\le L(u) + X_T\sqrt{\frac{2}{\beta} f(u)}\sqrt{M + \sum_{t\in\mathcal{U}} \frac{\eta_t^2 \|x_t\|_*^2 + 2\beta\eta_t y_t \langle w_t, x_t\rangle}{X_t^2}} - \sum_{t\in\mathcal{U}} \eta_t$$

where we have used the fact that $X_t \le X_T$ for all $t = 1, \ldots, T$. Solving for M we get

$$M \le L(u) + \frac{1}{\beta} f(u) X_T^2 + X_T\sqrt{\frac{2}{\beta} f(u)}\sqrt{\frac{1}{2\beta} X_T^2 f(u) + L(u) + D'} - \sum_{t\in\mathcal{U}} \eta_t$$

with $\frac{1}{2\beta} X_T^2 f(u) + L(u) + D' \ge 0$, and $D' = \sum_{t\in\mathcal{U}} \left(\frac{\eta_t^2 \|x_t\|_*^2 + 2\beta\eta_t y_t \langle w_t, x_t\rangle}{X_t^2} - \eta_t \right)$.

For the rest of proof, see the paper. □

Appendix

Given $(a_1, \ldots, a_d) \in \mathbb{R}_+$ and $q \in (1, 2]$, define the regularization function by

$$f(w) = \frac{1}{2(q-1)} \left(\sum_{i=1}^{d} |w_i|^q \, a_i \right)^{2/q}$$

Lemma 2

The Fenchel conjugate of f is

$$f^*(\theta) = \frac{1}{2(p-1)} \left(\sum_{i=1}^{d} |\theta_i|^p \, a_i^{1-p} \right)^{2/p} \quad \textit{for } p = \frac{q}{q-1}$$

Moreover, the function $f(w)$ is 1 -strongly convex with respect to the norm

$$\left(\sum_{i=1}^{d} |x_i|^q \, a_i \right)^{1/q}$$

whose dual norm is defined by

$$\left(\sum_{i=1}^{d} |\theta_i|^p \, a_i^{1-p} \right)^{1/p}.$$

Appendix

Lemma 3

Assume OMD is run with functions $f_1, f_2, \ldots, f_T$ *defined on* $\mathbb{X}$ *and such that each* f_t *is* β_t *strongly convex with respect to the norm* $\|\cdot\|_t$ *and* $f_t(\lambda u) \leq \lambda^2 f_t(u)$ *for all* $\lambda \in \mathbb{R}$ *and all* $u \in S$. *For each* $t = 1, 2, \ldots, T$ *let* $\|_{t,*}$ *be the dual norm of* $\|\cdot\|_t$. *Assume further the input sequence is* $z_t = -\eta_t \ell'_t$ *for some* $\eta_t > 0$, *where* $\ell'_t \in \partial \ell_t(w_t), \ell_t(w_t) = 0$ *implies* $\ell'_t = \mathbf{0}$, *and* $\ell_t = \ell(\langle \cdot, \boldsymbol{x}_t \rangle, y_t)$ *satisfies* (20). *Then, for all* $T \geq 1$

$$\sum_{t \in \mathcal{M} \cup \mathcal{U}} \eta_t \leq L_\eta + \lambda f_T(u) + \frac{1}{\lambda}\left(B + \sum_{t \in \mathcal{M} \cup \mathcal{U}} \left(\frac{\eta_t^2}{2\beta_t} \|\ell'_t\|_{t,*}^2 - \eta_t \langle w_t, \ell'_t \rangle\right)\right)$$

for any $u \in S$ *and any* $\lambda > 0$, *where*

$$L_\eta = \sum_{t \in \mathcal{M} \cup \mathcal{U}} \eta_t \ell_t(u) \text{ and } B = \sum_{t=1}^{T} \left(f_t^*(\theta_t) - f_{t-1}^*(\theta_t)\right)$$

In particular, choosing the optimal λ, *we obtain*

$$\sum_{t \in \mathcal{M} \cup \mathcal{U}} \eta_t \leq L_\eta + 2\sqrt{f_T(u)} \sqrt{\left[B + \sum_{t \in \mathcal{M} \cup U} \left(\frac{\eta_t^2}{2\beta_t} \|\ell'_t\|_{t,*}^2 - \eta_t \langle w_t, \ell'_t \rangle\right)\right]_+}.$$