Online Learning

— Online Mirror Descent (Part II)

Joseph Chuang-Chieh Lin

Department of Computer Science & Information Engineering,
Tamkang University

Spring 2023

Credits for the resource

The slides are based on the lectures of Prof. Luca Trevisan:
https://lucatrevisan.github.io/40391/index.html

the lectures of Prof. Shipra Agrawal:
https://ieor8100.github.io/mab/

the lectures of Prof. Francesco Orabona:
https://parameterfree.com/lecture-notes-on-online-learning/
the monograph: https://arxiv.org/abs/1912.13213

and also Elad Hazan's textbook:
Introduction to Online Convex Optimization, 2nd Edition.

Outline

1. A Short and Quick Review
2. The Mirror Interpretation
3. Another Way for the Update

Outline

Algorithm OMD

Input: Non-empty closed convex $V \subseteq X \subseteq \mathbb{R}^d$,
$\psi : X \mapsto \mathbb{R}$ strictly convex and continuously differentiable on $\mathrm{int}(X)$,
$\mathbf{x}_1 \in V$ s.t. ψ is differentiable in $\mathbf{x}_1$,
$\eta_1, \ldots, \eta_T > 0$.

1: **for** $t \leftarrow 1$ to T **do**
2: Output $\mathbf{x}_t$
3: Receive $f_t : \mathbb{R}^d \mapsto (-\infty, +\infty]$ and suffer $f_t(\mathbf{x}_t)$
4: Set $\mathbf{g}_t \in \partial f_t(\mathbf{x}_t)$
5: $\mathbf{x}_{t+1} \leftarrow \arg\min_{\mathbf{x} \in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{\eta_t} B_\psi(\mathbf{x}; \mathbf{x}_t)$
6: **end for**

Assume one of the following boundary conditions.

- $\lim_{\mathbf{x} \to \partial X} \|\nabla \psi(\mathbf{x})\|_2 = +\infty$.
- $V \subset \mathrm{int}(X)$.

Main Lemma

Lemma (Regret Inequality for OMD)

- ψ: λ-strongly convex w.r.t. $\|\cdot\|$ in V.
- B_ψ: the Bregman divergence w.r.t. $\psi : X \mapsto \mathbb{R}$.
- $V \subseteq X$: non-empty, closed & convex.
- Set $\mathbf{g}_t \in \partial f_t(\mathbf{x}_t)$.
- Assume one of the two boundary conditions holds.

Then for each $\mathbf{u} \in V$ and Algorithm OMD, we have

$$\eta_t(f_t(\mathbf{x}_t) - f_t(\mathbf{u})) \le \eta_t\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u}\rangle \le B_\psi(\mathbf{u}; \mathbf{x}_t) - B_\psi(\mathbf{u}; \mathbf{x}_{t+1}) + \frac{\eta_t^2}{2\lambda}\|\mathbf{g}_t\|_*^2.$$

Main Theorem

Main Theorem I

- Set $\mathbf{x}_1 \in V$ such that ψ is differentiable in $\mathbf{x}_1$.
- Assume that $\eta_{t+1} \leq \eta_t$ for $t = 1, \ldots, T$.

Then, under the assumption in the Main Lemma and $\forall \mathbf{u} \in V$, we have

$$\sum_{t=1}^{T} (f_t(\mathbf{x}_t) - f_t(\mathbf{u})) \leq \max_{1 \leq t \leq T} \frac{B_\psi(\mathbf{u}; \mathbf{x}_t)}{\eta_T} + \frac{1}{2\lambda} \sum_{t=1}^{T} \eta_t \|\mathbf{g}_t\|_*^2.$$

Remark on Main Theorem I

- The regret bound depends on arbitrary couple of dual norms $\|\cdot\|$ and $\|\cdot\|_*$.
  - Usually, the primal norm is used to measure the feasible set V or the distance between the competitor and the initial point.
  - The dual norm will be used to measure the gradients.

Outline

Why "Mirror"?

- Recall that $\mathbf{x} \in \partial f^*(\boldsymbol{\theta})$ if and only if $\boldsymbol{\theta} \in \partial f(\mathbf{x})$, for any closed and convex function $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$.

Why "Mirror"?

- Recall that $\mathbf{x} \in \partial f^*(\boldsymbol{\theta})$ if and only if $\boldsymbol{\theta} \in \partial f(\mathbf{x})$, for any closed and convex function $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$.
  - That is, $(\partial f)^{-1} = \partial f^*$.
- We will see that the Fenchel conjugate of a stronly convex function is smooth and differentiable.

Theorem (Duality Strong Convexity/Smoothness)

Let $\psi : \mathbb{R}^d \mapsto (-\infty, +\infty]$ be a closed function. Then ψ is λ-strongly convex w.r.t. to $\|\cdot\|$ if and only if

1. ψ^* is differentiable;
2. ψ^* is $\frac{1}{\lambda}$-smooth w.r.t. $\|\cdot\|_*$.

$(\Rightarrow)$:

- Since ψ is strongly convex, the maximizer $\mathbf{x}^*$ of $\max_{\mathbf{x}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \psi(\mathbf{x})$ exists and is unique!

Theorem (Duality Strong Convexity/Smoothness)

Let $\psi : \mathbb{R}^d \mapsto (-\infty, +\infty]$ be a closed function. Then ψ is λ-strongly convex w.r.t. to $\|\cdot\|$ if and only if

1. ψ^* is differentiable;
2. ψ^* is $\frac{1}{\lambda}$-smooth w.r.t. $\|\cdot\|_*$.

$(\Rightarrow)$:

- Since ψ is strongly convex, the maximizer $\mathbf{x}^*$ of $\max_{\mathbf{x}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \psi(\mathbf{x})$ exists and is unique!
- $\mathbf{x}^* \in \partial \psi^*(\boldsymbol{\theta})$.

Theorem (Duality Strong Convexity/Smoothness)

Let $\psi : \mathbb{R}^d \mapsto (-\infty, +\infty]$ be a closed function. Then ψ is λ-strongly convex w.r.t. to $\|\cdot\|$ if and only if

1. ψ^* is differentiable;
2. ψ^* is $\frac{1}{\lambda}$-smooth w.r.t. $\|\cdot\|_*$.

$(\Rightarrow)$:

- Since ψ is strongly convex, the maximizer $\mathbf{x}^*$ of $\max_{\mathbf{x}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \psi(\mathbf{x})$ exists and is unique!
- $\mathbf{x}^* \in \partial\psi^*(\boldsymbol{\theta})$.
- Suppose that $\exists \mathbf{x}' \in \partial\psi^*(\boldsymbol{\theta})$ and $\mathbf{x}' \neq \mathbf{x}$

Theorem (Duality Strong Convexity/Smoothness)

Let $\psi : \mathbb{R}^d \mapsto (-\infty, +\infty]$ be a closed function. Then ψ is λ-strongly convex w.r.t. to $\| \cdot \|$ if and only if

1. ψ^* is differentiable;
2. ψ^* is $\frac{1}{\lambda}$-smooth w.r.t. $\| \cdot \|_*$.

$(\Rightarrow)$:

- Since ψ is strongly convex, the maximizer $\mathbf{x}^*$ of $\max_{\mathbf{x}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \psi(\mathbf{x})$ exists and is unique!
- $\mathbf{x}^* \in \partial \psi^*(\boldsymbol{\theta})$.
- Suppose that $\exists \mathbf{x}' \in \partial \psi^*(\boldsymbol{\theta})$ and $\mathbf{x}' \neq \mathbf{x} \Rightarrow \psi^*(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{x}' \rangle - \psi(\mathbf{x}')$ Uniqueness implies that $\mathbf{x}' = \mathbf{x}$.

($\Rightarrow$):

Smoothness of ψ^*:

- For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, let $\mathbf{x}_1 = \nabla\psi^*(\boldsymbol{\theta}_1), \mathbf{x}_2 = \nabla\psi^*(\boldsymbol{\theta}_2)$.

(⇒):

Smoothness of ψ^*:

- For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, let $\mathbf{x}_1 = \nabla\psi^*(\boldsymbol{\theta}_1), \mathbf{x}_2 = \nabla\psi^*(\boldsymbol{\theta}_2)$.
- By the strong convexity of ψ, we have

$$\psi(\mathbf{x}_2) \geq \psi(\mathbf{x}_1) + \langle \boldsymbol{\theta}_1, \mathbf{x}_2 - \mathbf{x}_1 \rangle + \frac{\lambda}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2,$$
$$\psi(\mathbf{x}_1) \geq \psi(\mathbf{x}_2) + \langle \boldsymbol{\theta}_2, \mathbf{x}_1 - \mathbf{x}_2 \rangle + \frac{\lambda}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2.$$

Summing them we derive

$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_*\|\mathbf{x}_1 - \mathbf{x}_2\| \geq \langle \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1, \mathbf{x}_1 - \mathbf{x}_2 \rangle \geq \lambda\|\mathbf{x}_1 - \mathbf{x}_2\|^2$$

$(\Rightarrow)$:

Smoothness of ψ^*:

- For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, let $\mathbf{x}_1 = \nabla\psi^*(\boldsymbol{\theta_1}), \mathbf{x}_2 = \nabla\psi^*(\boldsymbol{\theta_2})$.
- By the strong convexity of ψ, we have

$$\psi(\mathbf{x}_2) \geq \psi(\mathbf{x}_1) + \langle \boldsymbol{\theta}_1, \mathbf{x}_2 - \mathbf{x}_1 \rangle + \frac{\lambda}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2,$$
$$\psi(\mathbf{x}_1) \geq \psi(\mathbf{x}_2) + \langle \boldsymbol{\theta}_2, \mathbf{x}_1 - \mathbf{x}_2 \rangle + \frac{\lambda}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2.$$

Summing them we derive

$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_* \|\mathbf{x}_1 - \mathbf{x}_2\| \geq \langle \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1, \mathbf{x}_1 - \mathbf{x}_2 \rangle \geq \lambda\|\mathbf{x}_1 - \mathbf{x}_2\|^2$$

So

$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_* \geq \lambda\|\mathbf{x}_1 - \mathbf{x}_2\| = \lambda\|\nabla\psi^*(\boldsymbol{\theta}_1) - \nabla\psi^*(\boldsymbol{\theta}_2)\|.$$

($\Leftarrow$):

Assume that ψ^* is differentiable and $(1/\lambda)$-smooth.

- Let $\mathbf{y} \in \mathrm{dom}(\psi)$ and $\mathbf{u} \in \partial\psi(\mathbf{y})$.
- By previous Lemma and the differentiability of ψ^*, we have $\mathbf{y} = \nabla\psi^*(\mathbf{u})$.

$(\Leftarrow)$:

Assume that ψ^* is differentiable and $(1/\lambda)$-smooth.

- Let $\mathbf{y} \in \text{dom}(\psi)$ and $\mathbf{u} \in \partial\psi(\mathbf{y})$.
- By previous Lemma and the differentiability of ψ^*, we have $\mathbf{y} = \nabla\psi^*(\mathbf{u})$.
- Define: $\phi(\boldsymbol{\theta}) := \psi^*(\boldsymbol{\theta} + \mathbf{u}) - \psi^*(\mathbf{u}) - \langle \nabla\psi^*(\mathbf{u}), \boldsymbol{\theta} \rangle$.

($\Leftarrow$):

Assume that ψ^* is differentiable and $(1/\lambda)$-smooth.

- Let $\mathbf{y} \in \mathrm{dom}(\psi)$ and $\mathbf{u} \in \partial\psi(\mathbf{y})$.
- By previous Lemma and the differentiability of ψ^*, we have $\mathbf{y} = \nabla\psi^*(\mathbf{u})$.
- Define: $\phi(\boldsymbol{\theta}) := \psi^*(\boldsymbol{\theta} + \mathbf{u}) - \psi^*(\mathbf{u}) - \langle \nabla\psi^*(\mathbf{u}), \boldsymbol{\theta} \rangle$.
- By Lemma 4.21 [in Prof. Orabona's Monograph] & the $1/\lambda$-smoothness of ψ^*, we have $\phi(\boldsymbol{\theta}) \leq \frac{1}{2\lambda}\|\boldsymbol{\theta}\|_*^2$ (Left as an exercise).

($\Leftarrow$):

Assume that ψ^* is differentiable and $(1/\lambda)$-smooth.

- Let $\mathbf{y} \in \mathrm{dom}(\psi)$ and $\mathbf{u} \in \partial\psi(\mathbf{y})$.
- By previous Lemma and the differentiability of ψ^*, we have $\mathbf{y} = \nabla\psi^*(\mathbf{u})$.
- Define: $\phi(\boldsymbol{\theta}) := \psi^*(\boldsymbol{\theta} + \mathbf{u}) - \psi^*(\mathbf{u}) - \langle \nabla\psi^*(\mathbf{u}), \boldsymbol{\theta} \rangle$.
- By Lemma 4.21 [in Prof. Orabona's Monograph] & the $1/\lambda$-smoothness of ψ^*, we have $\phi(\boldsymbol{\theta}) \leq \frac{1}{2\lambda}\|\boldsymbol{\theta}\|_*^2$ (Left as an exercise).
- Then we can obtain $\phi^*(\mathbf{x}) \geq \frac{\lambda}{2}\|\mathbf{x}\|^2$ (Left as an exercise).

($\Leftarrow$):

Calculate $\phi^*(\mathbf{x})$.

$$\begin{aligned}
\phi^*(\mathbf{x}) &= \sup_{\boldsymbol{\theta}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \psi^*(\boldsymbol{\theta} + \mathbf{u}) + \psi^*(\mathbf{u}) + \langle \boldsymbol{\theta}, \nabla \psi^*(\mathbf{u}) \rangle \\
&= \psi^*(\mathbf{u}) - \langle \mathbf{u}, \mathbf{x} + \nabla \psi^*(\mathbf{u}) \rangle + \sup_{\mathbf{v}} \langle \mathbf{v}, \mathbf{x} + \nabla \psi^*(\mathbf{u}) \rangle - \psi^*(\mathbf{v}) \\
&= \psi^*(\mathbf{u}) - \langle \mathbf{u}, \mathbf{x} + \nabla \psi^*(\mathbf{u}) \rangle + \psi^{**}(\mathbf{x} + \nabla \psi^*(\mathbf{u})) \\
&= \psi^*(\mathbf{u}) - \langle \mathbf{u}, \mathbf{x} + \nabla \psi^*(\mathbf{u}) \rangle + \psi(\mathbf{x} + \nabla \psi^*(\mathbf{u})) \\
&= -\langle \mathbf{u}, \mathbf{x} \rangle - \psi(\nabla \psi^*(\mathbf{u})) + \psi(\mathbf{x} + \nabla \psi^*(\mathbf{u})) \\
&= -\langle \mathbf{u}, \mathbf{x} \rangle - \psi(\mathbf{y}) + \psi(\mathbf{x} + \mathbf{y}).
\end{aligned}$$

- Note that $\langle \mathbf{u}, \nabla \psi^*(\mathbf{u}) \rangle = \psi^*(\mathbf{u}) + \psi(\nabla \psi^*(\mathbf{u}))$; let $\mathbf{v} := \boldsymbol{\theta} + \mathbf{u}$.

($\Leftarrow$):

Calculate $\phi^*(\mathbf{x})$.

$$\begin{aligned}
\phi^*(\mathbf{x}) &= \sup_{\boldsymbol{\theta}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \psi^*(\boldsymbol{\theta} + \mathbf{u}) + \psi^*(\mathbf{u}) + \langle \boldsymbol{\theta}, \nabla \psi^*(\mathbf{u}) \rangle \\
&= \psi^*(\mathbf{u}) - \langle \mathbf{u}, \mathbf{x} + \nabla \psi^*(\mathbf{u}) \rangle + \sup_{\mathbf{v}} \langle \mathbf{v}, \mathbf{x} + \nabla \psi^*(\mathbf{u}) \rangle - \psi^*(\mathbf{v}) \\
&= \psi^*(\mathbf{u}) - \langle \mathbf{u}, \mathbf{x} + \nabla \psi^*(\mathbf{u}) \rangle + \psi^{**}(\mathbf{x} + \nabla \psi^*(\mathbf{u})) \\
&= \psi^*(\mathbf{u}) - \langle \mathbf{u}, \mathbf{x} + \nabla \psi^*(\mathbf{u}) \rangle + \psi(\mathbf{x} + \nabla \psi^*(\mathbf{u})) \\
&= -\langle \mathbf{u}, \mathbf{x} \rangle - \psi(\nabla \psi^*(\mathbf{u})) + \psi(\mathbf{x} + \nabla \psi^*(\mathbf{u})) \\
&= -\langle \mathbf{u}, \mathbf{x} \rangle - \psi(\mathbf{y}) + \psi(\mathbf{x} + \mathbf{y}).
\end{aligned}$$

- Note that $\langle \mathbf{u}, \nabla \psi^*(\mathbf{u}) \rangle = \psi^*(\mathbf{u}) + \psi(\nabla \psi^*(\mathbf{u}))$; let $\mathbf{v} := \boldsymbol{\theta} + \mathbf{u}$.
- Thus, $\psi(\mathbf{x} + \mathbf{y}) - \psi(\mathbf{y}) - \langle \mathbf{u}, \mathbf{x} \rangle = \phi^*(\mathbf{x}) \geq \frac{\lambda}{2} \|\mathbf{x}\|^2$.

$(\Leftarrow)$:

Calculate $\phi^*(\mathbf{x})$.

$$
\begin{aligned}
\phi^*(\mathbf{x}) &= \sup_{\boldsymbol{\theta}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \psi^*(\boldsymbol{\theta} + \mathbf{u}) + \psi^*(\mathbf{u}) + \langle \boldsymbol{\theta}, \nabla \psi^*(\mathbf{u}) \rangle \\
&= \psi^*(\mathbf{u}) - \langle \mathbf{u}, \mathbf{x} + \nabla \psi^*(\mathbf{u}) \rangle + \sup_{\mathbf{v}} \langle \mathbf{v}, \mathbf{x} + \nabla \psi^*(\mathbf{u}) \rangle - \psi^*(\mathbf{v}) \\
&= \psi^*(\mathbf{u}) - \langle \mathbf{u}, \mathbf{x} + \nabla \psi^*(\mathbf{u}) \rangle + \psi^{**}(\mathbf{x} + \nabla \psi^*(\mathbf{u})) \\
&= \psi^*(\mathbf{u}) - \langle \mathbf{u}, \mathbf{x} + \nabla \psi^*(\mathbf{u}) \rangle + \psi(\mathbf{x} + \nabla \psi^*(\mathbf{u})) \\
&= -\langle \mathbf{u}, \mathbf{x} \rangle - \psi(\nabla \psi^*(\mathbf{u})) + \psi(\mathbf{x} + \nabla \psi^*(\mathbf{u})) \\
&= -\langle \mathbf{u}, \mathbf{x} \rangle - \psi(\mathbf{y}) + \psi(\mathbf{x} + \mathbf{y}).
\end{aligned}
$$

- Note that $\langle \mathbf{u}, \nabla \psi^*(\mathbf{u}) \rangle = \psi^*(\mathbf{u}) + \psi(\nabla \psi^*(\mathbf{u}))$; let $\mathbf{v} := \boldsymbol{\theta} + \mathbf{u}$.
- Thus, $\psi(\mathbf{x} + \mathbf{y}) - \psi(\mathbf{y}) - \langle \mathbf{u}, \mathbf{x} \rangle = \phi^*(\mathbf{x}) \geq \frac{\lambda}{2} \|\mathbf{x}\|^2$.

The First-Order Optimality Condition

Theorem (FO Optimality Condition)

Given $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$. Then $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ if and only if $\mathbf{0} \in \partial f(\mathbf{x}^*)$.

The First-Order Optimality Condition

Theorem (FO Optimality Condition)

Given $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$. Then $\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ if and only if $\mathbf{0} \in \partial f(\mathbf{x}^*)$.

$$\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \Leftrightarrow$$

The First-Order Optimality Condition

Theorem (FO Optimality Condition)

Given $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$. Then $\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ if and only if $\mathbf{0} \in \partial f(\mathbf{x}^*)$.

$$\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \Leftrightarrow \forall \mathbf{y} \in \mathbb{R}^d, f(\mathbf{y}) \geq f(\mathbf{x}^*)$$

The First-Order Optimality Condition

Theorem (FO Optimality Condition)

Given $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$. Then $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ if and only if $\mathbf{0} \in \partial f(\mathbf{x}^*)$.

$$\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \Leftrightarrow \forall \mathbf{y} \in \mathbb{R}^d, f(\mathbf{y}) \geq f(\mathbf{x}^*) = f(\mathbf{x}^*) + \langle \mathbf{0}, \mathbf{y} - \mathbf{x}^* \rangle$$

The First-Order Optimality Condition

Theorem (FO Optimality Condition)

Given $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$. Then $\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ if and only if $\mathbf{0} \in \partial f(\mathbf{x}^*)$.

$$\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \Leftrightarrow \forall \mathbf{y} \in \mathbb{R}^d, f(\mathbf{y}) \geq f(\mathbf{x}^*) = f(\mathbf{x}^*) + \langle \mathbf{0}, \mathbf{y} - \mathbf{x}^* \rangle \Leftrightarrow \mathbf{0} \in \partial f(\mathbf{x}^*).$$

The Mirror & Bregman Divergence

Theorem (Mirror & Bregman Divergence)

Let B_ψ be the Bregman divergence w.r.t. a λ-strongly convex and closed $\psi : X \mapsto \mathbb{R}$, where $\lambda > 0$. Let $V \subseteq X$ be a non-empty closed convex set and $\mathbf{x}_t \in V$. Define

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{\eta_t} B_\psi(\mathbf{x}; \mathbf{x}_t).$$

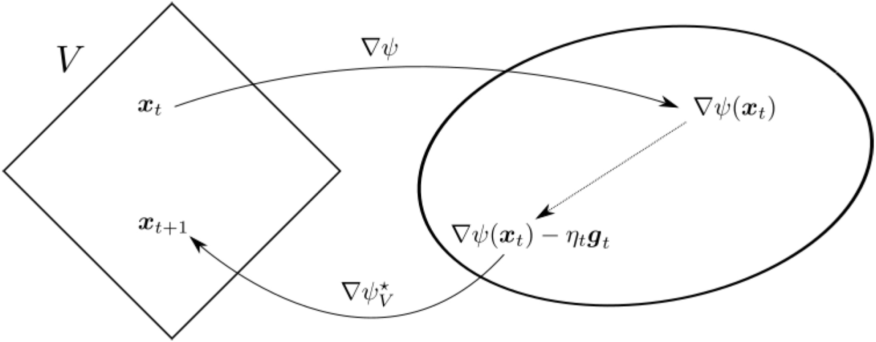
Assume ψ to be differentiable in $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$. Then, for any $\mathbf{g}_t \in \mathbf{R}^d$, we have

$$\mathbf{x}_{t+1} = \nabla \psi_V^*(\nabla \phi(\mathbf{x}_t) - \eta_t \mathbf{g}_t),$$

where $\psi_V := \psi + i_V$ is the restriction of ψ to V, and

$$i_V(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in V \\ 1 & \text{otherwise} \end{cases}$$

Illustration (refer to Prof. Orabona's Monograph, Ch.6)



The Proof (1/2)

$$
\begin{aligned}
\mathbf{x}_{t+1} &= \operatorname*{arg\,min}_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{\eta_t} B_\psi(\mathbf{x}; \mathbf{x}_t) \\
&= \operatorname*{arg\,min}_{\mathbf{x}\in V} \eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + B_\psi(\mathbf{x}; \mathbf{x}_t) \\
&= \operatorname*{arg\,min}_{\mathbf{x}\in V} \eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle \\
&= \operatorname*{arg\,min}_{\mathbf{x}\in V} \langle \eta_t \mathbf{g}_t - \nabla \psi(\mathbf{x}_t), \mathbf{x} \rangle + \psi(\mathbf{x}).
\end{aligned}
$$

The Proof (1/2)

$$
\begin{aligned}
\mathbf{x}_{t+1} &= \arg\min_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x}\rangle + \frac{1}{\eta_t} B_\psi(\mathbf{x}; \mathbf{x}_t) \\
&= \arg\min_{\mathbf{x}\in V} \eta_t \langle \mathbf{g}_t, \mathbf{x}\rangle + B_\psi(\mathbf{x}; \mathbf{x}_t) \\
&= \arg\min_{\mathbf{x}\in V} \eta_t \langle \mathbf{g}_t, \mathbf{x}\rangle + \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t\rangle \\
&= \arg\min_{\mathbf{x}\in V} \langle \eta_t \mathbf{g}_t - \nabla\psi(\mathbf{x}_t), \mathbf{x}\rangle + \psi(\mathbf{x}).
\end{aligned}
$$

Use the first-order optimality condition:

$$
\mathbf{0} \in \eta_t \mathbf{g}_t + \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t) + \partial i_V(\mathbf{x}_{t+1}).
$$

The Proof (1/2)

$$
\begin{aligned}
\mathbf{x}_{t+1} &= \arg\min_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{\eta_t} B_\psi(\mathbf{x}; \mathbf{x}_t) \\
&= \arg\min_{\mathbf{x}\in V} \eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + B_\psi(\mathbf{x}; \mathbf{x}_t) \\
&= \arg\min_{\mathbf{x}\in V} \eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle \\
&= \arg\min_{\mathbf{x}\in V} \langle \eta_t \mathbf{g}_t - \nabla\psi(\mathbf{x}_t), \mathbf{x} \rangle + \psi(\mathbf{x}).
\end{aligned}
$$

Use the first-order optimality condition:

$$
\mathbf{0} \in \eta_t \mathbf{g}_t + \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t) + \partial i_V(\mathbf{x}_{t+1}).
$$

$\Rightarrow$

$$
\nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t \in \nabla\psi(\mathbf{x}_{t+1}) + \partial i_V(\mathbf{x}_{t+1}) \subseteq \partial\psi_V(\mathbf{x}_{t+1}).
$$

The Proof (2/2)

$$\nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t \in \partial\psi_V(\mathbf{x}_{t+1})$$

implies

The Proof (2/2)

$$\nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t \in \partial\psi_V(\mathbf{x}_{t+1})$$

implies

$$\mathbf{x}_{t+1} \in \partial\psi_V^*(\nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t).$$

The Proof (2/2)

$$\nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t \in \partial\psi_V(\mathbf{x}_{t+1})$$

implies

$$\mathbf{x}_{t+1} \in \partial\psi_V^*(\nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t).$$

Since ψ_V is λ-strongly convex and closed, we have $\partial\psi_V^* = \{\nabla\psi_V^*\}$.

Therefore,

The Proof (2/2)

$$\nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t \in \partial\psi_V(\mathbf{x}_{t+1})$$

implies

$$\mathbf{x}_{t+1} \in \partial\psi_V^*(\nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t).$$

Since ψ_V is λ-strongly convex and closed, we have $\partial\psi_V^* = \{\nabla\psi_V^*\}$.

Therefore,

$$\mathbf{x}_{t+1} = \nabla\psi_V^*(\nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t).$$

The Reasons...

- OMD extends the OSD to non-Euclidean norms.
  - Dual norms can be considered.
- It makes sense to use a dual norm to measure a gradient.
  - How "big" the linear functional $\mathbf{x} \mapsto \langle f(\mathbf{y}), \mathbf{x} \rangle$ is.

The Reasons...

- OMD extends the OSD to non-Euclidean norms.
  - Dual norms can be considered.
- It makes sense to use a dual norm to measure a gradient.
  - How "big" the linear functional $\mathbf{x} \mapsto \langle f(\mathbf{y}), \mathbf{x} \rangle$ is.
- Gradients actually live in the *dual space*, which is different from where the predictions $\mathbf{x}_t$ live

The Reasons...

- OMD extends the OSD to non-Euclidean norms.
  - Dual norms can be considered.
- It makes sense to use a dual norm to measure a gradient.
  - How "big" the linear functional $\mathbf{x} \mapsto \langle f(\mathbf{y}), \mathbf{x} \rangle$ is.
- Gradients actually live in the *dual space*, which is different from where the predictions $\mathbf{x}_t$ live
- So, why do we apply OSD??

Example for $\psi(\mathbf{x})$ being the L_2-Norm (1/3)

- Let $\psi : \mathbf{R}^d \mapsto \mathbb{R}$, $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$.
- $V = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$.
- Let $\psi_V := \psi + i_V$.

Example for $\psi(\mathbf{x})$ being the L_2-Norm (1/3)

- Let $\psi : \mathbf{R}^d \mapsto \mathbb{R}$, $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$.
- $V = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$.
- Let $\psi_V := \psi + i_V$.
- Then,

$$\psi_V^*(\boldsymbol{\theta}) =$$

Example for $\psi(\mathbf{x})$ being the L_2-Norm (1/3)

- Let $\psi : \mathbf{R}^d \mapsto \mathbb{R}$, $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$.
- $V = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$.
- Let $\psi_V := \psi + i_V$.
- Then,

$$\psi_V^*(\boldsymbol{\theta}) = \sup_{\mathbf{x} \in V} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \frac{1}{2}\|\mathbf{x}\|_2^2.$$

Example for $\psi(\mathbf{x})$ being the L_2-Norm (1/3)

- Let $\psi : \mathbf{R}^d \mapsto \mathbb{R}$, $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$.
- $V = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$.
- Let $\psi_V := \psi + i_V$.
- Then,

$$\psi_V^*(\boldsymbol{\theta}) = \sup_{\mathbf{x} \in V} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \frac{1}{2}\|\mathbf{x}\|_2^2.$$

- First, note that $\psi_V^*(\boldsymbol{\theta}) = \mathbf{0}$ if $\boldsymbol{\theta} = \mathbf{0}$.

Example for $\psi(\mathbf{x})$ being the L_2-Norm (2/3)

- For any $\mathbf{x} \in V$, there exist $\mathbf{q}$ and $\alpha \in \mathbb{R}$ such that
  - $\mathbf{x} = \alpha \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2} + \mathbf{q}$.
  - $\langle \mathbf{q}, \boldsymbol{\theta} \rangle = 0$.
- So, we have

$$
\begin{aligned}
& \sup_{\mathbf{x} \in V} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \frac{1}{2}\|\mathbf{x}\|_2^2 \\
= & \sup_{\substack{\alpha, \mathbf{q}: \\ \alpha \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2} + \mathbf{q} \in V, \langle \mathbf{q}, \boldsymbol{\theta} \rangle = 0}} \alpha \|\boldsymbol{\theta}\|_2 - \frac{\alpha^2}{2} - \frac{1}{2}\|\mathbf{q}\|_2^2 \\
= & \sup_{-1 \le \alpha \le 1} \alpha \|\boldsymbol{\theta}\|_2 - \frac{\alpha^2}{2}.
\end{aligned}
$$

Example for $\psi(\mathbf{x})$ being the L_2-Norm (3/3)

- Solving the above optimization problem, we have

$$\psi_V^*(\boldsymbol{\theta}) = \begin{cases} \frac{1}{2}\|\boldsymbol{\theta}\|_2^2, & \|\boldsymbol{\theta}\|_2 \leq 1 \\ \|\boldsymbol{\theta}\|_2 - \frac{1}{2}, & \|\boldsymbol{\theta}\|_2 > 1 \end{cases},$$

  which is finite everywhere and differentiable.

Example for $\psi(\mathbf{x})$ being the L_2-Norm (3/3)

- Solving the above optimization problem, we have

$$\psi_V^*(\boldsymbol{\theta}) = \begin{cases} \frac{1}{2}\|\boldsymbol{\theta}\|_2^2, & \|\boldsymbol{\theta}\|_2 \leq 1 \\ \|\boldsymbol{\theta}\|_2 - \frac{1}{2}, & \|\boldsymbol{\theta}\|_2 > 1 \end{cases},$$

which is finite everywhere and differentiable.

- Therefore, we have $\nabla\psi(\mathbf{x}) = \mathbf{x}$, and

$$\nabla\psi_V^*(\boldsymbol{\theta}) = \begin{cases} \boldsymbol{\theta}, & \|\boldsymbol{\theta}\|_2 \leq 1 \\ \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2}, & \|\boldsymbol{\theta}\|_2 > 1 \end{cases} = \Pi_V(\boldsymbol{\theta}).$$

Example for $\psi(\mathbf{x})$ being the L_2-Norm (3/3)

- Solving the above optimization problem, we have

$$\psi_V^*(\boldsymbol{\theta}) = \begin{cases} \frac{1}{2}\|\boldsymbol{\theta}\|_2^2, & \|\boldsymbol{\theta}\|_2 \leq 1 \\ \|\boldsymbol{\theta}\|_2 - \frac{1}{2}, & \|\boldsymbol{\theta}\|_2 > 1 \end{cases},$$

which is finite everywhere and differentiable.

- Therefore, we have $\nabla\psi(\mathbf{x}) = \mathbf{x}$, and

$$\nabla\psi_V^*(\boldsymbol{\theta}) = \begin{cases} \boldsymbol{\theta}, & \|\boldsymbol{\theta}\|_2 \leq 1 \\ \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2}, & \|\boldsymbol{\theta}\|_2 > 1 \end{cases} = \Pi_V(\boldsymbol{\theta}).$$

⋆ This is exactly the update of projected online subgradient descent.

Outline

An Equivalent Two-Step Update (1/2)

Theorem (Two-Step Update)

- Let $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$ be closed, strictly convex and differentiable in int dom(f).
- Let $V \subset \mathbb{R}^d$ be a non-empty, closed and convex set, such that $V \cap (\mathrm{f}) \neq \emptyset$.
- Assume that $\tilde{\mathbf{y}} = \arg\min_{\mathbf{z} \in \mathbb{R}^d} f(\mathbf{z})$ exists and $\tilde{\mathbf{y}} \in \mathrm{int\ dom}(f)$.
- Denote by $\mathbf{y}' = \arg\min_{\mathbf{z} \in V} B_f(\mathbf{z}; \tilde{\mathbf{y}})$.

Then the following hold:

1. $\mathbf{y} = \arg\min_{\mathbf{z} \in V} f(\mathbf{z})$ exists and is unique.
2. $\mathbf{y} = \mathbf{y}'$.

An Equivalent Two-Step Update (2/2)

Therefore, under the assumption of the theorem, we have that

$$\mathbf{x}_{t+1} = \operatorname*{arg\,min}_{\mathbf{x} \in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{\eta_t} B_\psi(\mathbf{x}; \mathbf{x}_t)$$

is equivalent to

$$\tilde{\mathbf{x}}_{t+1} = \operatorname*{arg\,min}_{\mathbf{x} \in \mathbb{R}^d} \langle \eta_t \mathbf{g}_t, \mathbf{x} \rangle + B_\psi(\mathbf{x}; \mathbf{x}_t)$$

$$\mathbf{x}_{t+1} = \operatorname*{arg\,min}_{\mathbf{x} \in V} B_\psi(\mathbf{x}; \tilde{\mathbf{x}}_{t+1})$$

Discussions