Online Learning

— Online Mirror Descent

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.

I would like to especially thank Prof. Francesco Orabona for the discussion with me about the details for this part of lectures.

Outline

Outline

Online Subgradient Descent (OSD)

- Consider the simplified case that $f_t(\cdot) = f(\cdot)$ for all $t > 0$.
- The key property for the convergence of OSD:

$$f(\mathbf{x}_t) - f(\mathbf{u}) \leq \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle, \ \forall \mathbf{u}.$$

Online Subgradient Descent (OSD)

- Consider the simplified case that $f_t(\cdot) = f(\cdot)$ for all $t > 0$.
- The key property for the convergence of OSD:

$$f(\mathbf{x}_t) - f(\mathbf{u}) \leq \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle, \ \forall \mathbf{u}.$$

- However, for $\mathbf{x} \in \mathbb{R}^2$, consider the following two functions:
  - $f(\mathbf{x}) = \max\{-x_1, x_1 - x_2, x_1 + x_2\}$.
  - $f(\mathbf{x}) = \max\{x_1^2 + (x_2 + 1)^2, x_1^2 + (x_2 - 1)^2\}$.

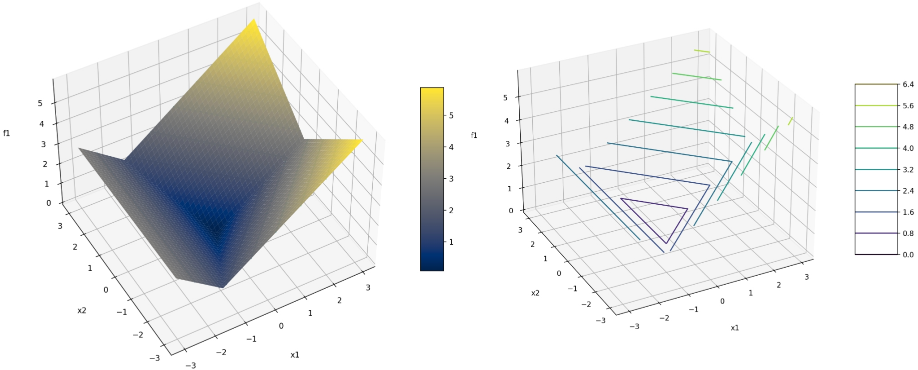
Online Subgradient Descent (OSD)

- Consider the simplified case that $f_t(\cdot) = f(\cdot)$ for all $t > 0$.
- The key property for the convergence of OSD:

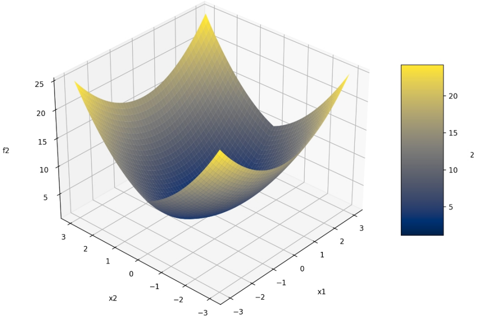
$$f(\mathbf{x}_t) - f(\mathbf{u}) \le \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle, \ \forall \mathbf{u}.$$

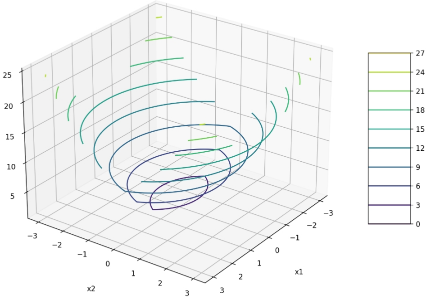
- However, for $\mathbf{x} \in \mathbb{R}^2$, consider the following two functions:
  - $f(\mathbf{x}) = \max\{-x_1, x_1 - x_2, x_1 + x_2\}$.
  - $f(\mathbf{x}) = \max\{x_1^2 + (x_2+1)^2, x_1^2 + (x_2-1)^2\}$.
- Moving toward the direction of the negative subgradient may not decrease the objective (loss).

Uninformative Subgradients



Uninformative Subgradients





Outline

A Linear Lower Bound by a Subgradient

- We can have a linear lower bound on function f around $\mathbf{x}_0$:

$$f(\mathbf{x}) \geq \tilde{f}(\mathbf{x}) := f(\mathbf{x}_0) + \langle \mathbf{g}, \mathbf{x} - \mathbf{x_0} \rangle, \forall \mathbf{x} \in V.$$

  - Let's say $V \subseteq \mathbb{R}^d$ is the domain.

A Linear Lower Bound by a Subgradient

- We can have a linear lower bound on function f around $\mathbf{x}_0$:

$$f(\mathbf{x}) \geq \tilde{f}(\mathbf{x}) := f(\mathbf{x}_0) + \langle \mathbf{g}, \mathbf{x} - \mathbf{x_0} \rangle, \; \forall \mathbf{x} \in V.$$

  - Let's say $V \subseteq \mathbb{R}^d$ is the domain.
- Note that over unbounded domains the minimizer of linear function at the right-hand side above is $-\infty$.

A Principle of Moderation

- Minimizing the previous lower bound only in a neighborhood of $\mathbf{x}_0$.

$$
\begin{aligned}
\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in V} f(\mathbf{x}_t) + \langle \mathbf{g}, \mathbf{x} - \mathbf{x}_t \rangle \\
\text{subject to } \|\mathbf{x}_t - \mathbf{x}\|^2 \leq h, \\
\text{for some } h > 0.
\end{aligned}
$$

A Principle of Moderation

- Minimizing the previous lower bound only in a neighborhood of $\mathbf{x}_0$.

$$\begin{aligned} \mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in V} & f(\mathbf{x}_t) + \langle \mathbf{g}, \mathbf{x} - \mathbf{x}_t \rangle \\ \text{subject to } & \|\mathbf{x}_t - \mathbf{x}\|^2 \leq h, \\ & \text{for some } h > 0. \end{aligned}$$

- Unconstrained formulation: (assume $\eta > 0$)

$$\arg\min_{\mathbf{x} \in V} f(\mathbf{x}_0) + \langle \mathbf{g}, \mathbf{x} - \mathbf{x}_0 \rangle + \frac{1}{2\eta} \|\mathbf{x}_0 - \mathbf{x}\|_2^2.$$

Solving the minimization (ignore non-variable terms):

$$\arg\min_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{2\eta_t}\|\mathbf{x}_t - \mathbf{x}\|_2^2 = \arg\min_{\mathbf{x}\in V} 2\eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + \|\mathbf{x}_t - \mathbf{x}\|_2^2$$

Solving the minimization (ignore non-variable terms):

$$\begin{aligned}
&\arg\min_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x}\rangle + \frac{1}{2\eta_t}\|\mathbf{x}_t - \mathbf{x}\|_2^2 = \arg\min_{\mathbf{x}\in V} 2\eta_t\langle \mathbf{g}_t, \mathbf{x}\rangle + \|\mathbf{x}_t - \mathbf{x}\|_2^2 \\
= \quad &\arg\min_{\mathbf{x}\in V} \|\eta_t \mathbf{g}_t\|_2^2 + 2\eta_t\langle \mathbf{g}_t, \mathbf{x} - \mathbf{x}_t\rangle + \|\mathbf{x}_t - \mathbf{x}\|_2^2
\end{aligned}$$

Solving the minimization (ignore non-variable terms):

$$
\begin{aligned}
& \arg\min_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{2\eta_t}\|\mathbf{x}_t - \mathbf{x}\|_2^2 = \arg\min_{\mathbf{x}\in V} 2\eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + \|\mathbf{x}_t - \mathbf{x}\|_2^2 \\
= \quad & \arg\min_{\mathbf{x}\in V} \|\eta_t \mathbf{g}_t\|_2^2 + 2\eta_t \langle \mathbf{g}_t, \mathbf{x} - \mathbf{x}_t \rangle + \|\mathbf{x}_t - \mathbf{x}\|_2^2 \\
= \quad & \arg\min_{\mathbf{x}\in V} \|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{x}\|_2^2
\end{aligned}
$$

Solving the minimization (ignore non-variable terms):

$$
\begin{aligned}
& \arg\min_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}\|_2^2 = \arg\min_{\mathbf{x}\in V} 2\eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + \|\mathbf{x}_t - \mathbf{x}\|_2^2 \\
= \quad & \arg\min_{\mathbf{x}\in V} \|\eta_t \mathbf{g}_t\|_2^2 + 2\eta_t \langle \mathbf{g}_t, \mathbf{x} - \mathbf{x}_t \rangle + \|\mathbf{x}_t - \mathbf{x}\|_2^2 \\
= \quad & \arg\min_{\mathbf{x}\in V} \|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{x}\|_2^2 \\
= \quad & \Pi_V(\mathbf{x}_t - \eta_t \mathbf{g}_t),
\end{aligned}
$$

where $\Pi_V(\mathbf{x}) = \arg\min_{\mathbf{y}\in V} \|\mathbf{x} - \mathbf{y}\|_2$ (Euclidean projection onto V).

Solving the minimization (ignore non-variable terms):

$$
\begin{aligned}
& \arg\min_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}\|_2^2 = \arg\min_{\mathbf{x}\in V} 2\eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + \|\mathbf{x}_t - \mathbf{x}\|_2^2 \\
= \quad & \arg\min_{\mathbf{x}\in V} \|\eta_t \mathbf{g}_t\|_2^2 + 2\eta_t \langle \mathbf{g}_t, \mathbf{x} - \mathbf{x}_t \rangle + \|\mathbf{x}_t - \mathbf{x}\|_2^2 \\
= \quad & \arg\min_{\mathbf{x}\in V} \|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{x}\|_2^2 \\
= \quad & \Pi_V(\mathbf{x}_t - \eta_t \mathbf{g}_t),
\end{aligned}
$$

where $\Pi_V(\mathbf{x}) = \arg\min_{\mathbf{y}\in V} \|\mathbf{x} - \mathbf{y}\|_2$ (Euclidean projection onto V).

- So, we rediscovered the online subgradient descent with projection!

The Inspiration

- Choosing a different norm or distance measure for the locality of $\mathbf{x}$ leads to a different updating method.

The Inspiration

- Choosing a different norm or distance measure for the locality of $\mathbf{x}$ leads to a different updating method.
- Change to which norm?

The Inspiration

- Choosing a different norm or distance measure for the locality of $\mathbf{x}$ leads to a different updating method.
- Change to which norm?
- Any alternative to norms?

The Inspiration

- Choosing a different norm or distance measure for the locality of $\mathbf{x}$ leads to a different updating method.
- Change to which norm?
- Any alternative to norms?

$$\arg\min_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}\|_2^2$$

The Inspiration

- Choosing a different norm or distance measure for the locality of $\mathbf{x}$ leads to a different updating method.
- Change to which norm?
- Any alternative to norms?

$$
\begin{aligned}
& \arg\min_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}\|_2^2 \\
\Rightarrow \quad & \arg\min_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{\eta_t} B_\psi(\mathbf{x}; \mathbf{x}_t)
\end{aligned}
$$

The Inspiration

- Choosing a different norm or distance measure for the locality of $\mathbf{x}$ leads to a different updating method.
- Change to which norm?
- Any alternative to norms?

$$
\begin{aligned}
& \arg\min_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{2\eta_t}\|\mathbf{x}_t - \mathbf{x}\|_2^2 \\
\Rightarrow \quad & \arg\min_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{\eta_t} B_\psi(\mathbf{x}; \mathbf{x}_t)
\end{aligned}
$$

Note: When $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, the two updates are exactly the same.

Outline

Strictly Convexity

Strictly Convex Functions

A function $f : V \subseteq \mathbb{R}^d \mapsto \mathbb{R}$, where V is a convex set, is strictly convex if

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) < \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}),$$

$\forall \mathbf{x}, \mathbf{y} \in V, \mathbf{x} \neq \mathbf{y}, \alpha \in (0, 1)$.

- Strong convexity w.r.t. any norm implies strict convexity.

Strictly Convexity

Strictly Convex Functions

A function $f : V \subseteq \mathbb{R}^d \mapsto \mathbb{R}$, where V is a convex set, is strictly convex if

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) < \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}),$$

$\forall \mathbf{x}, \mathbf{y} \in V, \mathbf{x} \neq \mathbf{y}, \alpha \in (0,1)$.

- Strong convexity w.r.t. any norm implies strict convexity.
- If f is differentiable, strict convexity implies that

$$f(\mathbf{y}) > f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

for $\mathbf{x} \neq \mathbf{y}$.

Bregman Divergence

Bregman Divergence

Let $\psi : X \mapsto \mathbb{R}$ be strictly convex and continuously differentiable on $\text{int}(X)$. The Bregman Divergence w.r.t. ψ is $B_\psi : X \times \text{int}(X) \mapsto \mathbb{R}$ defined as

$$B_\psi(\mathbf{x}; \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

- Always non-negative ($\because \psi$ is convex).

Bregman Divergence

Bregman Divergence

Let $\psi : X \mapsto \mathbb{R}$ be strictly convex and continuously differentiable on $\text{int}(X)$. The Bregman Divergence w.r.t. ψ is $B_\psi : X \times \text{int}(X) \mapsto \mathbb{R}$ defined as

$$B_\psi(\mathbf{x}; \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

- Always non-negative ($\because$ ψ is convex).
- $\psi(\mathbf{x}) \geq \psi(\mathbf{y}) + \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle, \forall \mathbf{y} \in X,$

Bregman Divergence

Bregman Divergence

Let $\psi : X \mapsto \mathbb{R}$ be strictly convex and continuously differentiable on $\text{int}(X)$. The Bregman Divergence w.r.t. ψ is $B_\psi : X \times \text{int}(X) \mapsto \mathbb{R}$ defined as

$$B_\psi(\mathbf{x}; \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

- Always non-negative ($\because \psi$ is convex).
- $\psi(\mathbf{x}) \geq \psi(\mathbf{y}) + \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle, \forall \mathbf{y} \in X$, and equality holds only for $\mathbf{y} = \mathbf{x}$

Bregman Divergence

Bregman Divergence

Let $\psi : X \mapsto \mathbb{R}$ be strictly convex and continuously differentiable on $\mathrm{int}(X)$. The Bregman Divergence w.r.t. ψ is $B_\psi : X \times \mathrm{int}(X) \mapsto \mathbb{R}$ defined as

$$B_\psi(\mathbf{x}; \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

- Always non-negative ($\because \psi$ is convex).
- $\psi(\mathbf{x}) \geq \psi(\mathbf{y}) + \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle, \forall \mathbf{y} \in X$, and equality holds only for $\mathbf{y} = \mathbf{x}$ ($\because$ strictly convexity of ψ).

Bregman Divergence

Bregman Divergence

Let $\psi : X \mapsto \mathbb{R}$ be strictly convex and continuously differentiable on $\text{int}(X)$. The Bregman Divergence w.r.t. ψ is $B_\psi : X \times \text{int}(X) \mapsto \mathbb{R}$ defined as

$$B_\psi(\mathbf{x}; \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

- Always non-negative ($\because$ ψ is convex).
- $\psi(\mathbf{x}) \geq \psi(\mathbf{y}) + \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle, \forall \mathbf{y} \in X$, and equality holds only for $\mathbf{y} = \mathbf{x}$ ($\because$ strictly convexity of ψ).
- It can be a distance measure, though it is NOT Symmetric.

Examples (1/5)

- Consider a twice differentiable ψ in a ball B around $\mathbf{y}$ and $\mathbf{x} \in B$.
- By Taylor's theorem, there exists $0 \leq \alpha \leq 1$ such that

$$B_\psi(\mathbf{x}; \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \nabla\psi(\mathbf{y})^\top(\mathbf{x} - \mathbf{y})$$

Examples (1/5)

- Consider a twice differentiable ψ in a ball B around $\mathbf{y}$ and $\mathbf{x} \in B$.
- By Taylor's theorem, there exists $0 \leq \alpha \leq 1$ such that

$$B_\psi(\mathbf{x}; \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \nabla\psi(\mathbf{y})^\top(\mathbf{x} - \mathbf{y}) = \frac{1}{2}(\mathbf{x} - \mathbf{y})^\top \nabla^2\psi(\mathbf{z})(\mathbf{x} - \mathbf{y}),$$

for $\mathbf{z} = \alpha\mathbf{x} + (1 - \alpha)\mathbf{y}$.

Examples (1/5)

- Consider a twice differentiable ψ in a ball B around $\mathbf{y}$ and $\mathbf{x} \in B$.
- By Taylor's theorem, there exists $0 \le \alpha \le 1$ such that

$$B_\psi(\mathbf{x}; \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \nabla\psi(\mathbf{y})^\top(\mathbf{x} - \mathbf{y}) = \frac{1}{2}(\mathbf{x} - \mathbf{y})^\top \nabla^2\psi(\mathbf{z})(\mathbf{x} - \mathbf{y}),$$

for $\mathbf{z} = \alpha\mathbf{x} + (1 - \alpha)\mathbf{y}$.

⋆ A squared local norm depending on the Hessian of ψ.

Examples (2/5)

- If ψ is λ-**strongly convex** w.r.t. a norm $\|\cdot\|$ in $\mathrm{int}(X)$,

Examples (2/5)

- If ψ is λ-**strongly convex** w.r.t. a norm $\|\cdot\|$ in $\text{int}(X)$, we have $B_\psi(\mathbf{x};\mathbf{y}) \geq \frac{\lambda}{2}\|\mathbf{x}-\mathbf{y}\|^2$.

Examples (3/5)

- If $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, then

$$B_\psi(\mathbf{x}; \mathbf{y}) = \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{y}\|_2^2 -$$

Examples (3/5)

- If $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, then

$$B_\psi(\mathbf{x};\mathbf{y}) = \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{y}\|_2^2 - \langle \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle$$

Examples (3/5)

- If $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, then

$$B_\psi(\mathbf{x}; \mathbf{y}) = \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{y}\|_2^2 - \langle \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2.$$

Examples (4/5): **Exercise**

Please show that:

- If $\psi(\mathbf{x}) = \sum_{i=1}^{d} x_i \ln x_i$, and $X = \{\mathbf{x} \mid x_i \geq 0, \|\mathbf{x}\|_1 = 1\}$, then

$$B_\psi(\mathbf{x}; \mathbf{y}) = \sum_{i=1}^{d} x_i \ln \frac{x_i}{y_i}.$$

Examples (4/5): **Exercise**

Please show that:

- If $\psi(\mathbf{x}) = \sum_{i=1}^{d} x_i \ln x_i$, and $X = \{\mathbf{x} \mid x_i \geq 0, \|\mathbf{x}\|_1 = 1\}$, then

$$B_\psi(\mathbf{x}; \mathbf{y}) = \sum_{i=1}^{d} x_i \ln \frac{x_i}{y_i}.$$

⋆ This is the Kullback-Leibler divergence (KL-divergence) between two distributions **x** and **y**.

Examples (5/5): **Exercise**

Please prove the following lemma.

Lemma [Chen & Teboulle 1993]

Let B_ψ be the Bregman divergence w.r.t. $\psi : X \mapsto \mathbb{R}$. Then, for any three points $\mathbf{x}, \mathbf{y} \in \mathrm{int}(X)$ and $\mathbf{z} \in X$, we have

$$B_\psi(\mathbf{z}; \mathbf{x}) + B_\psi(\mathbf{x}; \mathbf{y}) - B_\psi(\mathbf{z}; \mathbf{y}) = \langle \nabla\psi(\mathbf{y}) - \nabla\psi(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle.$$

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 $\mathbf{x}_1$,
$\eta_1, \dots, \eta_T > 0$.

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**

Fix Some Minor Issues

Add one of the following boundary conditions.

- $\lim_{\lambda \to 0} \langle \nabla \psi(\mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle = -\infty$, for any $\mathbf{x} \in \text{boundary}(X)$, $\mathbf{y} \in \text{int}(X)$.
- $V \subseteq \text{int}(X)$.

When $\arg\min$ exists, $\mathbf{x}_{t+1} \in \operatorname{int}(X)$

Theorem

Let

- B_ψ be the Bregman divergence w.r.t. $\psi : X \mapsto \mathbb{R}$.
- $V \subseteq X$ be a non-empty closed and convex set.

Assume that previous two boundary conditions holds and the $\arg\min$ of the algorithm exists on all rounds, then we have $\mathbf{x}_{t+1} \in \operatorname{int}(X)$.

Existence of the $\arg\min$'s

Theorem

Let

- $\lambda > 0$
- $f : \mathbb{R} \mapsto (-\infty, +\infty]$ a closed and λ-strongly convex w.r.t. $\|\cdot\|$.

Assume that $\mathrm{dom}(\partial f) \neq \emptyset$. Then, f has exactly one minimizer.

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

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

Proof of the Main Lemma (1/3)

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**

Proof of the Main Lemma (1/3)

Input: Non-empty closed convex $V \subseteq X \subseteq \mathbb{R}^d$,
$\psi : X \mapsto \mathbb{R}$ strictly convex and continuously differentiable on $\text{int}(X)$,
$\mathbf{x}_1 \in V$ 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**

$$\frac{\partial}{\partial \mathbf{x}} \left(\eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + B_\psi(\mathbf{x}; \mathbf{x}_t)\right)$$

Proof of the Main Lemma (1/3)

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**

$$\frac{\partial}{\partial \mathbf{x}} \left(\eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + B_\psi(\mathbf{x}; \mathbf{x}_t)\right) = \eta_t \mathbf{g}_t + \nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{x}_t)$$

Proof of the Main Lemma (1/3)

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

```
for t ← 1 to T do
    Output x_t
    Receive f_t : R^d ↦ (−∞, +∞] and suffer f_t(x_t)
    Set g_t ∈ ∂f_t(x_t)
    x_{t+1} ← argmin_{x∈V} ⟨g_t, x⟩ + (1/η_t) B_ψ(x; x_t)
end for
```

$$\frac{\partial}{\partial \mathbf{x}} \left(\eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + B_\psi(\mathbf{x}; \mathbf{x}_t) \right) = \eta_t \mathbf{g}_t + \nabla \psi(\mathbf{x}) - \nabla \psi(\mathbf{x}_t)$$

The optimality condition guarantees that

Proof of the Main Lemma (1/3)

Input: Non-empty closed convex $V \subseteq X \subseteq \mathbb{R}^d$,
$\psi : X \mapsto \mathbb{R}$ strictly convex and continuously differentiable on $\text{int}(X)$,
$\mathbf{x}_1 \in V$ 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**

$$\frac{\partial}{\partial \mathbf{x}} \left(\eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + B_\psi(\mathbf{x}; \mathbf{x}_t)\right) = \eta_t \mathbf{g}_t + \nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{x}_t)$$

The optimality condition guarantees that

$$\langle \eta_t \mathbf{g}_t + \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1} \rangle \geq 0, \ \forall \mathbf{u} \in V.$$

Proof of the Main Lemma (2/3)

$$
\begin{aligned}
\langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle &= -\langle \eta_t \mathbf{g}_t + \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1} \rangle \\
&\quad + \langle \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1} \rangle + \langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\
&\leq \langle \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1} \rangle + \langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\
&= B_\psi(\mathbf{u}; \mathbf{x}_t) - B_\psi(\mathbf{u}; \mathbf{x}_{t+1}) - B_\psi(\mathbf{x}_{t+1}; \mathbf{x}_t) + \langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\
&\leq B_\psi(\mathbf{u}; \mathbf{x}_t) - B_\psi(\mathbf{u}; \mathbf{x}_{t+1}) - \frac{\lambda}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \eta_t \|\mathbf{g}_t\|_* \|\mathbf{x}_t - \mathbf{x}_{t+1}\| \\
&\leq B_\psi(\mathbf{u}; \mathbf{x}_t) - B_\psi(\mathbf{u}; \mathbf{x}_{t+1}) + \frac{\eta_t^2}{2\lambda}\|\mathbf{g}_t\|_*^2.
\end{aligned}
$$

Proof of the Main Lemma (2/3)

$$
\begin{aligned}
\langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle &= -\langle \eta_t \mathbf{g}_t + \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1} \rangle \\
&\quad + \langle \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1} \rangle + \langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\
&\leq \langle \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1} \rangle + \langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\
&= B_\psi(\mathbf{u}; \mathbf{x}_t) - B_\psi(\mathbf{u}; \mathbf{x}_{t+1}) - B_\psi(\mathbf{x}_{t+1}; \mathbf{x}_t) + \langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\
&\leq B_\psi(\mathbf{u}; \mathbf{x}_t) - B_\psi(\mathbf{u}; \mathbf{x}_{t+1}) - \frac{\lambda}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \eta_t \|\mathbf{g}_t\|_* \|\mathbf{x}_t - \mathbf{x}_{t+1}\| \\
&\leq B_\psi(\mathbf{u}; \mathbf{x}_t) - B_\psi(\mathbf{u}; \mathbf{x}_{t+1}) + \frac{\eta_t^2}{2\lambda} \|\mathbf{g}_t\|_*^2.
\end{aligned}
$$

Proof of the Main Lemma (2/3)

$$
\begin{aligned}
\langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle &= -\langle \eta_t \mathbf{g}_t + \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1} \rangle \\
&\quad + \langle \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1} \rangle + \langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\
&\leq \langle \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1} \rangle + \langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\
&= B_\psi(\mathbf{u}; \mathbf{x}_t) - B_\psi(\mathbf{u}; \mathbf{x}_{t+1}) - B_\psi(\mathbf{x}_{t+1}; \mathbf{x}_t) + \langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\
&\leq B_\psi(\mathbf{u}; \mathbf{x}_t) - B_\psi(\mathbf{u}; \mathbf{x}_{t+1}) - \frac{\lambda}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \eta_t \|\mathbf{g}_t\|_* \|\mathbf{x}_t - \mathbf{x}_{t+1}\| \\
&\leq B_\psi(\mathbf{u}; \mathbf{x}_t) - B_\psi(\mathbf{u}; \mathbf{x}_{t+1}) + \frac{\eta_t^2}{2\lambda} \|\mathbf{g}_t\|_*^2 .
\end{aligned}
$$

Hint

$$ax - \frac{b}{2}x^2 \leq \frac{a^2}{2b}, \text{ for } x \in \mathbb{R} \text{ and } a, b > 0.$$

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

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

Proof of Main Theorem I

$$
\begin{aligned}
\sum_{t=1}^{T}(f_t(\mathbf{x}_t) - f_t(\mathbf{u})) &\le \sum_{t=1}^{T}\left(\frac{1}{\eta_t}B_\psi(\mathbf{u};\mathbf{x}_t) - \frac{1}{\eta_t}B_\psi(\mathbf{u};\mathbf{x}_{t+1})\right) + \sum_{t=1}^{T}\frac{\eta_t^2}{2\lambda}\|\mathbf{g}_t\|_*^2 \\
&= \frac{1}{\eta_1}B_\psi(\mathbf{u};\mathbf{x}_1) - \frac{1}{\eta_T}B_\psi(\mathbf{u};\mathbf{x}_{T+1}) + \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)B_\psi(\mathbf{u};\mathbf{x}_{t+1}) + \sum_{t=1}^{T}\frac{\eta_t}{2\lambda}\|\mathbf{g}_t\|_*^2 \\
&\le \frac{1}{\eta_1}D^2 + D^2\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) + \sum_{t=1}^{T}\frac{\eta_t}{2\lambda}\|\mathbf{g}_t\|_*^2 \\
&= \frac{1}{\eta_1}D^2 + D^2\left(\frac{1}{\eta_T} - \frac{1}{\eta_1}\right) + \sum_{t=1}^{T}\frac{\eta_t}{2\lambda}\|\mathbf{g}_t\|_*^2 \\
&= \frac{D^2}{\eta_T} + \sum_{t=1}^{T}\frac{\eta_t}{2\lambda}\|\mathbf{g}_t\|_*^2,
\end{aligned}
$$

where $D^2 := \max_{1\le t\le T} B_\psi(\mathbf{u};\mathbf{x}_t)$.

What we can learn from OMD?

- OMD allows us to prove regret guarantees depending on arbitrary norms $\|\cdot\|$ and $\|\cdot\|_*$.
- The primal norm: measure in the feasible space.
- The dual norm: measuring the gradients.

Using a Fixed Learning Rate $\eta_t = \eta$

- Assume that f_t is L-Lipschitz continuous $\Rightarrow \|\mathbf{g}_t\|_*^2 = \|\mathbf{g}_t\|_2^2 \leq L^2$.
- To minimize $\frac{D^2}{\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2\lambda}\|\mathbf{g}_t\|_*^2 = \frac{D^2}{\eta} + \frac{T\eta L^2}{2\lambda}$.
  - Take the derivative w.r.t. η and get root: $\Rightarrow \frac{D^2}{\eta^2} = \frac{TL^2}{2\lambda}$, $\eta = \frac{\sqrt{2\lambda}D}{L\sqrt{T}}$
  - Then the regret is $\frac{DL\sqrt{2T}}{\sqrt{\lambda}}$.

Using Adaptive Learning Rate

- Set $\eta_t = \frac{D\sqrt{\lambda}}{\sqrt{\sum_{i=1}^{t} \|\mathbf{g}_i\|_2^2}}$.
- We can show that

$$\sum_{t=1}^{T} \frac{\eta_t}{2\lambda} \|\mathbf{g}_t\|_2^2 = \frac{D}{2\sqrt{\lambda}} \sum_{t=1}^{T} \frac{\|\mathbf{g}_t\|_2^2}{\sum_{i=1}^{t} \|\mathbf{g}_i\|_2^2}$$

Using Adaptive Learning Rate

- Set $\eta_t = \frac{D\sqrt{\lambda}}{\sqrt{\sum_{i=1}^{t} \|\mathbf{g}_i\|_2^2}}$.
- We can show that

$$\sum_{t=1}^{T} \frac{\eta_t}{2\lambda} \|\mathbf{g}_t\|_2^2 = \frac{D}{2\sqrt{\lambda}} \sum_{t=1}^{T} \frac{\|\mathbf{g}_t\|_2^2}{\sum_{i=1}^{t} \|\mathbf{g}_i\|_2^2} \leq \frac{D}{2\sqrt{\lambda}} \cdot 2\sqrt{\sum_{t=1}^{T} \|\mathbf{g}_t\|_2^2}$$

Using Adaptive Learning Rate

- Set $\eta_t = \frac{D\sqrt{\lambda}}{\sqrt{\sum_{i=1}^{t} \|\mathbf{g}_i\|_2^2}}$.
- We can show that

$$\sum_{t=1}^{T} \frac{\eta_t}{2\lambda} \|\mathbf{g}_t\|_2^2 = \frac{D}{2\sqrt{\lambda}} \sum_{t=1}^{T} \frac{\|\mathbf{g}_t\|_2^2}{\sum_{i=1}^{t} \|\mathbf{g}_i\|_2^2} \leq \frac{D}{2\sqrt{\lambda}} \cdot 2\sqrt{\sum_{t=1}^{T} \|\mathbf{g}_t\|_2^2} \leq \frac{DL\sqrt{T}}{\sqrt{\lambda}}$$

Using Adaptive Learning Rate

- Set $\eta_t = \frac{D\sqrt{\lambda}}{\sqrt{\sum_{i=1}^{t} \|\mathbf{g}_i\|_2^2}}$.
- We can show that

$$\sum_{t=1}^{T} \frac{\eta_t}{2\lambda} \|\mathbf{g}_t\|_2^2 = \frac{D}{2\sqrt{\lambda}} \sum_{t=1}^{T} \frac{\|\mathbf{g}_t\|_2^2}{\sum_{i=1}^{t} \|\mathbf{g}_i\|_2^2} \leq \frac{D}{2\sqrt{\lambda}} \cdot 2\sqrt{\sum_{t=1}^{T} \|\mathbf{g}_t\|_2^2} \leq \frac{DL\sqrt{T}}{\sqrt{\lambda}}$$

- The regret turns out to be $\leq \frac{D^2}{\eta_T} + \frac{DL\sqrt{T}}{\sqrt{\lambda}}$, which is bounded by

Using Adaptive Learning Rate

- Set $\eta_t = \frac{D\sqrt{\lambda}}{\sqrt{\sum_{i=1}^{t} \|\mathbf{g}_i\|_2^2}}$.
- We can show that

$$\sum_{t=1}^{T} \frac{\eta_t}{2\lambda} \|\mathbf{g}_t\|_2^2 = \frac{D}{2\sqrt{\lambda}} \sum_{t=1}^{T} \frac{\|\mathbf{g}_t\|_2^2}{\sum_{i=1}^{t} \|\mathbf{g}_i\|_2^2} \leq \frac{D}{2\sqrt{\lambda}} \cdot 2\sqrt{\sum_{t=1}^{T} \|\mathbf{g}_t\|_2^2} \leq \frac{DL\sqrt{T}}{\sqrt{\lambda}}$$

- The regret turns out to be $\leq \frac{D^2}{\eta_T} + \frac{DL\sqrt{T}}{\sqrt{\lambda}}$, which is bounded by $\frac{2D}{\sqrt{\lambda}} \sqrt{\sum_{t=1}^{T} \|\mathbf{g}_t\|_2^2}$

Using Adaptive Learning Rate

- Set $\eta_t = \frac{D\sqrt{\lambda}}{\sqrt{\sum_{i=1}^{t} \|\mathbf{g}_i\|_2^2}}$.
- We can show that

$$\sum_{t=1}^{T} \frac{\eta_t}{2\lambda} \|\mathbf{g}_t\|_2^2 = \frac{D}{2\sqrt{\lambda}} \sum_{t=1}^{T} \frac{\|\mathbf{g}_t\|_2^2}{\sum_{i=1}^{t} \|\mathbf{g}_i\|_2^2} \leq \frac{D}{2\sqrt{\lambda}} \cdot 2\sqrt{\sum_{t=1}^{T} \|\mathbf{g}_t\|_2^2} \leq \frac{DL\sqrt{T}}{\sqrt{\lambda}}$$

- The regret turns out to be $\leq \frac{D^2}{\eta_T} + \frac{DL\sqrt{T}}{\sqrt{\lambda}}$, which is bounded by $\frac{2D}{\sqrt{\lambda}} \sqrt{\sum_{t=1}^{T} \|\mathbf{g}_t\|_2^2} \leq \frac{2DL\sqrt{T}}{\sqrt{\lambda}}$.

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

Duality Strong Convexity/Smoothness

Theorem

Let

- $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$ be a closed and convex function
- $\mathrm{dom}(\partial f) \neq \emptyset$

Then for $\lambda > 0$, f is λ-strongly convex w.r.t. $\|\cdot\|$ iff f^* is $\frac{1}{\lambda}$-smooth w.r.t. $\|\cdot\|_*$ on $\mathbb{R}^d$.

- f^* is differentiable.

Duality Strong Convexity/Smoothness

Theorem

Let

- $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$ be a closed and convex function
- $\mathrm{dom}(\partial f) \neq \emptyset$

Then for $\lambda > 0$, f is λ-strongly convex w.r.t. $\|\cdot\|$ iff f^* is $\frac{1}{\lambda}$-smooth w.r.t. $\|\cdot\|_*$ on $\mathbb{R}^d$.

- f^* is differentiable.
  - f is proper, closed and **strongly** convex $\Rightarrow$ the maximizer $\mathbf{x}^*$ of $\max_{\mathbf{x}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - f(\mathbf{x})$ exists and is **unique** (p. 24).

Duality Strong Convexity/Smoothness

Theorem

Let

- $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$ be a closed and convex function
- $\text{dom}(\partial f) \neq \emptyset$

Then for $\lambda > 0$, f is λ-strongly convex w.r.t. $\|\cdot\|$ iff f^* is $\frac{1}{\lambda}$-smooth w.r.t. $\|\cdot\|_*$ on $\mathbb{R}^d$.

- f^* is differentiable.
  - f is proper, closed and **strongly** convex $\Rightarrow$ the maximizer $\mathbf{x}^*$ of $\max_{\mathbf{x}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - f(\mathbf{x})$ exists and is **unique** (p. 24).
  - Hence, $\mathbf{x}^* \in \partial f^*(\boldsymbol{\theta})$.

Duality Strong Convexity/Smoothness

Theorem

Let

- $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$ be a closed and convex function
- $\text{dom}(\partial f) \neq \emptyset$

Then for $\lambda > 0$, f is λ-strongly convex w.r.t. $\|\cdot\|$ iff f^* is $\frac{1}{\lambda}$-smooth w.r.t. $\|\cdot\|_*$ on $\mathbb{R}^d$.

- f^* is differentiable.
  - f is proper, closed and **strongly** convex $\Rightarrow$ the maximizer $\mathbf{x}^*$ of $\max_{\mathbf{x}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - f(\mathbf{x})$ exists and is **unique** (p. 24).
  - Hence, $\mathbf{x}^* \in \partial f^*(\boldsymbol{\theta})$.
  - Assume another $\mathbf{x}' \in \partial f^*(\boldsymbol{\theta}) \Rightarrow f^*(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{x}' \rangle - f(\mathbf{x}')$.
  - By the uniqueness of the maximizer, we have $\mathbf{x}^* = \mathbf{x}'$.

($\Rightarrow$ contd.)

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

$$
\begin{aligned}
f(\mathbf{x}_2) &\geq f(\mathbf{x}_1) + \langle \boldsymbol{\theta}_1, \mathbf{x}_2 - \mathbf{x}_1 \rangle + \frac{\lambda}{2} \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \\
f(\mathbf{x}_1) &\geq f(\mathbf{x}_2) + \langle \boldsymbol{\theta}_2, \mathbf{x}_1 - \mathbf{x}_2 \rangle + \frac{\lambda}{2} \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \\
\Rightarrow \quad & \langle \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1, \mathbf{x}_1 - \mathbf{x}_2 \rangle \geq \lambda \|\mathbf{x}_1 - \mathbf{x}_2\|^2.
\end{aligned}
$$

($\Rightarrow$ contd.)

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

$$\begin{aligned} f(\mathbf{x}_2) &\geq f(\mathbf{x}_1) + \langle \boldsymbol{\theta}_1, \mathbf{x}_2 - \mathbf{x}_1 \rangle + \frac{\lambda}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2 \\ f(\mathbf{x}_1) &\geq f(\mathbf{x}_2) + \langle \boldsymbol{\theta}_2, \mathbf{x}_1 - \mathbf{x}_2 \rangle + \frac{\lambda}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2 \end{aligned}$$

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

($\Leftarrow$)

- Assume that f^* is $\frac{1}{\lambda}$-smooth w.r.t. $\|\cdot\|_*$ on $\mathbb{R}^d$.
- Let $\mathbf{y} \in \mathrm{dom}(\partial f)$ and $\mathbf{u} \in \partial f(\mathbf{y})$.
- Since f^* is differentiable, we have $\mathbf{y} = \nabla f^*(\mathbf{u})$.

$(\Leftarrow)$

- Assume that f^* is $\frac{1}{\lambda}$-smooth w.r.t. $\|\cdot\|_*$ on $\mathbb{R}^d$.
- Let $\mathbf{y} \in \mathrm{dom}(\partial f)$ and $\mathbf{u} \in \partial f(\mathbf{y})$.
- Since f^* is differentiable, we have $\mathbf{y} = \nabla f^*(\mathbf{u})$.
- Define $\phi(\boldsymbol{\theta}) := f^*(\boldsymbol{\theta} + \mathbf{u}) - f^*(\mathbf{u}) - \langle \boldsymbol{\theta}, \nabla f^*(\mathbf{u}) \rangle$.

Recall that if $f : V \mapsto \mathbb{R}$ is M-smooth, then for any $\mathbf{x}, \mathbf{y} \in V$ we have

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \frac{M}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

($\Leftarrow$)

- Assume that f^* is $\frac{1}{\lambda}$-smooth w.r.t. $\|\cdot\|_*$ on $\mathbb{R}^d$.
- Let $\mathbf{y} \in \mathrm{dom}(\partial f)$ and $\mathbf{u} \in \partial f(\mathbf{y})$.
- Since f^* is differentiable, we have $\mathbf{y} = \nabla f^*(\mathbf{u})$.
- Define $\phi(\boldsymbol{\theta}) := f^*(\boldsymbol{\theta} + \mathbf{u}) - f^*(\mathbf{u}) - \langle \boldsymbol{\theta}, \nabla f^*(\mathbf{u}) \rangle$.

Recall that if $f : V \mapsto \mathbb{R}$ is M-smooth, then for any $\mathbf{x}, \mathbf{y} \in V$ we have

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \frac{M}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

- Hence, $\phi(\boldsymbol{\theta}) \leq \frac{1}{2\lambda} \|\boldsymbol{\theta}\|_*^2 := \hat{\phi}(\theta)$.

($\Leftarrow$)

- Assume that f^* is $\frac{1}{\lambda}$-smooth w.r.t. $\|\cdot\|_*$ on $\mathbb{R}^d$.
- Let $\mathbf{y} \in \mathrm{dom}(\partial f)$ and $\mathbf{u} \in \partial f(\mathbf{y})$.
- Since f^* is differentiable, we have $\mathbf{y} = \nabla f^*(\mathbf{u})$.
- Define $\phi(\boldsymbol{\theta}) := f^*(\boldsymbol{\theta} + \mathbf{u}) - f^*(\mathbf{u}) - \langle \boldsymbol{\theta}, \nabla f^*(\mathbf{u}) \rangle$.

Recall that if $f : V \mapsto \mathbb{R}$ is M-smooth, then for any $\mathbf{x}, \mathbf{y} \in V$ we have

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \frac{M}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

- Hence, $\phi(\boldsymbol{\theta}) \leq \frac{1}{2\lambda} \|\boldsymbol{\theta}\|_*^2 := \hat{\phi}(\theta)$.

$$\phi^*(\mathbf{x}) \geq \hat{\phi}^*(\mathbf{x}) = \sup_{\boldsymbol{\theta}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \frac{1}{2\lambda} \|\boldsymbol{\theta}\|_*^2$$

($\Leftarrow$)

- Assume that f^* is $\frac{1}{\lambda}$-smooth w.r.t. $\|\cdot\|_*$ on $\mathbb{R}^d$.
- Let $\mathbf{y} \in \mathrm{dom}(\partial f)$ and $\mathbf{u} \in \partial f(\mathbf{y})$.
- Since f^* is differentiable, we have $\mathbf{y} = \nabla f^*(\mathbf{u})$.
- Define $\phi(\boldsymbol{\theta}) := f^*(\boldsymbol{\theta} + \mathbf{u}) - f^*(\mathbf{u}) - \langle \boldsymbol{\theta}, \nabla f^*(\mathbf{u}) \rangle$.

Recall that if $f : V \mapsto \mathbb{R}$ is M-smooth, then for any $\mathbf{x}, \mathbf{y} \in V$ we have

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \le \frac{M}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

- Hence, $\phi(\boldsymbol{\theta}) \le \frac{1}{2\lambda} \|\boldsymbol{\theta}\|_*^2 := \hat{\phi}(\theta)$.

$$\begin{aligned}
\phi^*(\mathbf{x}) \ge \hat{\phi}^*(\mathbf{x}) &= \sup_{\boldsymbol{\theta}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \frac{1}{2\lambda} \|\boldsymbol{\theta}\|_*^2 \le \sup_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_* \|\mathbf{x}\| - \frac{1}{2\lambda} \|\boldsymbol{\theta}\|_*^2 \\
&= \sup_{\boldsymbol{\theta}} \left(-\frac{1}{2\lambda} (\|\boldsymbol{\theta}\|_*^2 - 2\lambda \|\mathbf{x}\| \|\boldsymbol{\theta}\|_* + (\lambda \|\mathbf{x}\|)^2) \right) \\
&= \frac{\lambda}{2} \|\mathbf{x}\|^2.
\end{aligned}$$

($\Leftarrow$) Recall that $\phi(\boldsymbol{\theta}) := f^*(\boldsymbol{\theta}+\mathbf{u}) - f^*(\mathbf{u}) - \langle \boldsymbol{\theta}, \nabla f^*(\mathbf{u})\rangle$.

- Calculate $\phi^*(\mathbf{x})$: (Let $\mathbf{v} = \boldsymbol{\theta} + \mathbf{u}$)

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

- Using $\phi^*(\mathbf{x}) \geq \frac{\lambda}{2}\|\mathbf{x}\|^2$ then we are done.

First-order optimality

Theorem

For a function $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$, we have

$$\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \text{ if and only if } \mathbf{0} \in \partial f(\mathbf{x}^*).$$

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

The OMD update in terms of duality mappings

Theorem (OMD & Duality Mappings)

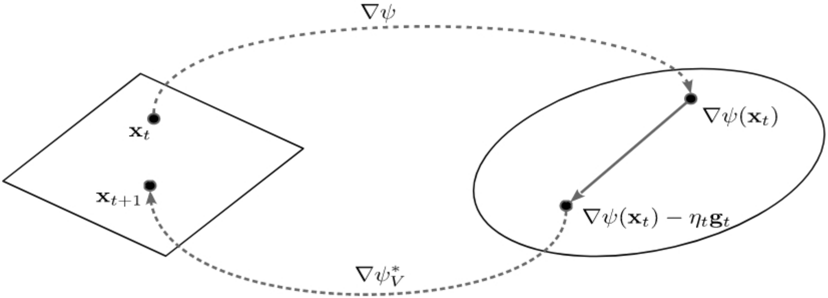
- Let B_ψ be the Bregman divergence w.r.t. $\psi : X \mapsto \mathbb{R}$, where ψ is closed and λ-strongly convex for $\lambda > 0$.
- Define $\mathbf{x}_{t+1} := \arg\min_{\mathbf{x}\in V} \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{\eta_t} B_\psi(\mathbf{x}; \mathbf{x}_t)$, and assume that ψ is differentiable in $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$.

Then, for any $\mathbf{g}_t \in \mathbb{R}^d$, we have

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

where $\psi_V := \psi + i_V$ which restricts ψ to V.

The OMD update in terms of duality mappings



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

Proof of the main theorem (1/2)

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

By the first-order optimality condition, we have

$$
\begin{aligned}
\mathbf{0} &\in \eta_t \mathbf{g}_t + \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t) + \partial i_V(\mathbf{x}_{t+1}) \\
\nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t &\in (\nabla\psi + \partial i_V)(\mathbf{x}_{t+1}) \subseteq \partial\psi_V(\mathbf{x}_{t+1})
\end{aligned}
$$

Proof of the main theorem (1/2)

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

By the first-order optimality condition, we have

$$
\begin{aligned}
\mathbf{0} &\in \eta_t \mathbf{g}_t + \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t) + \partial i_V(\mathbf{x}_{t+1}) \\
\nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t &\in (\nabla\psi + \partial i_V)(\mathbf{x}_{t+1}) \subseteq \partial\psi_V(\mathbf{x}_{t+1})
\end{aligned}
$$

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

Proof of the main theorem (2/2)

- Note that $\psi_V := \psi + i_V$ is proper, λ-strongly convex and closed.
  - $\partial\psi_V^* = \{\nabla\psi_V^*\}$.
- Therefore, since $\mathbf{x}_{t+1} \in \partial\psi_V^*(\nabla\psi(\mathbf{x}_t) - \eta_t\mathbf{g}_t)$, we have $\mathbf{x}_{t+1} = \nabla\psi_V^*(\nabla\psi(\mathbf{x}_t) - \eta_t\mathbf{g}_t)$.

Example (1/2)

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

Example (1/2)

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

- $\psi_V^*(\boldsymbol{\theta}) = \sup_{\mathbf{x} \in V} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \frac{1}{2}\|\mathbf{x}\|_2^2$.
- Assume $\boldsymbol{\theta} \neq \mathbf{0}$ (otherwise, trivially $\psi_V^*(\boldsymbol{\theta}) = \mathbf{0}$).

Example (1/2)

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

- $\psi_V^*(\boldsymbol{\theta}) = \sup_{\mathbf{x}\in V}\langle \boldsymbol{\theta}, \mathbf{x}\rangle - \frac{1}{2}\|\mathbf{x}\|_2^2$.
- Assume $\boldsymbol{\theta} \neq \mathbf{0}$ (otherwise, trivially $\psi_V^*(\boldsymbol{\theta}) = \mathbf{0}$).
- For any $\mathbf{x} \in V$, there exists $\mathbf{q}$ and α such that $\mathbf{x} = \alpha\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2} + \mathbf{q}$ and $\langle \mathbf{q}, \boldsymbol{\theta}\rangle = 0$.
    - $\nabla\psi(\mathbf{x}) = \mathbf{x}$.

Example (2/2)

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

- Solving the constrained optimization problem, we have $\alpha^*=\min(1,\|\boldsymbol{\theta}\|_2)$.
- Hence,

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

Example (2/2)

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

- Solving the constrained optimization problem, we have $\alpha^*=\min(1,\|\boldsymbol{\theta}\|_2)$.
- Hence,

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

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

Remark

- OMD extends the OSD to non-Euclidean norms.
- The dual norm is used to measure a gradient.
- Gradients live in the dual space, different from the predictions $\mathbf{x}_t$.
- In the OSD, the dual space coincides with the primal space.
- The ways we go from one space to the other: $\nabla\psi$ and $\nabla\psi_V^*$.

Discussions