Online Learning

— Fenchel Conjugate (Prerequisites for Mirror Descent Algorithms)

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

Outline

1. Fenchel Conjugate

2. Properties & Examples

Closed Function

Closed Function

A function $f : V \subseteq \mathbb{R}^d \mapsto [-\infty, +\infty]$ is closed if and only if $\{\mathbf{x} : f(\mathbf{x}) \leq \alpha\}$ is closed for every $\alpha \in \mathbb{R}$.

Closed Function

Closed Function

A function $f : V \subseteq \mathbb{R}^d \mapsto [-\infty, +\infty]$ is closed if and only if $\{\mathbf{x} : f(\mathbf{x}) \le \alpha\}$ is closed for every $\alpha \in \mathbb{R}$.

- $I_V : \mathbb{R}^d \mapsto \{0, 1\}$ is closed if and only if V is closed.

$$I_V(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in V \\ \infty & \text{if } \mathbf{x} \notin V \end{cases}$$

Fenchel Conjugate

- A generalization of Lagrangian duality which applies to non-convex functions.

Fenchel conjugate

Given a function $f : \mathbb{R}^d \mapsto [-\infty, +\infty]$, the Fenchel conjugate (or convex conjugate) $f^* : \mathbb{R}^d \mapsto [-\infty, +\infty]$ is defined as

$$f^*(\boldsymbol{\theta}) = \sup_{\mathbf{x} \in \mathbb{R}^d} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - f(\mathbf{x}).$$

Fenchel Conjugate

- A generalization of Lagrangian duality which applies to non-convex functions.

Fenchel conjugate

Given a function $f : \mathbb{R}^d \mapsto [-\infty, +\infty]$, the Fenchel conjugate (or convex conjugate) $f^* : \mathbb{R}^d \mapsto [-\infty, +\infty]$ is defined as

$$f^*(\boldsymbol{\theta}) = \sup_{\mathbf{x} \in \mathbb{R}^d} \langle \boldsymbol{\theta}, \mathbf{x} \rangle - f(\mathbf{x}).$$

Fenchel-Young's Inequality

$$\langle \boldsymbol{\theta}, \mathbf{x} \rangle \leq f(\mathbf{x}) + f^*(\boldsymbol{\theta}), \text{ for any } \mathbf{x}, \boldsymbol{\theta} \in \mathbb{R}^d.$$

Theorem [Rockafellar 1970]

Let f be a convex function. Then f^* is a closed convex function. Moreover, if f is closed, then $f^{**} = f$.

- A very detailed video of lecture as a reference: [link1], [link2].

Outline

Important Properties

Theorem

Let $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$ be a convex function. Then the following conditions are equivalent:

1. $\boldsymbol{\theta} \in \partial f(\mathbf{x})$.
2. $\langle \boldsymbol{\theta}, \mathbf{y} \rangle - f(\mathbf{y})$ achieves its supremum in $\mathbf{y}$ at $\mathbf{y} = \mathbf{x}$.
3. $f(\mathbf{x}) + f^*(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{x} \rangle$.
4. (If f is closed) $\mathbf{x} \in \partial f^*(\boldsymbol{\theta})$.

Proof (Sketch)

- (1) $\Leftrightarrow$ (2):
  - From the definition of subgradient:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \boldsymbol{\theta}, \mathbf{y} - \mathbf{x} \rangle, \ \forall \mathbf{y}.$$

Proof (Sketch)

- (1) $\Leftrightarrow$ (2):
  - From the definition of subgradient:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \boldsymbol{\theta}, \mathbf{y} - \mathbf{x} \rangle, \ \forall \mathbf{y}.$$

$$\Rightarrow \langle \boldsymbol{\theta}, \mathbf{x} \rangle - f(\mathbf{x}) \geq \langle \boldsymbol{\theta}, \mathbf{y} \rangle - f(\mathbf{y}), \ \forall \mathbf{y}.$$

Proof (Sketch)

- (1) $\Leftrightarrow$ (2):
  - From the definition of subgradient:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \boldsymbol{\theta}, \mathbf{y} - \mathbf{x} \rangle, \ \forall \mathbf{y}.$$

$$\Rightarrow \langle \boldsymbol{\theta}, \mathbf{x} \rangle - f(\mathbf{x}) \geq \langle \boldsymbol{\theta}, \mathbf{y} \rangle - f(\mathbf{y}), \ \forall \mathbf{y}.$$

- (2) $\Leftrightarrow$ (3): by definition of $f^*(\boldsymbol{\theta})$.

Proof (Sketch)

- (1) $\Leftrightarrow$ (2):
  - From the definition of subgradient:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \boldsymbol{\theta}, \mathbf{y} - \mathbf{x} \rangle, \ \forall \mathbf{y}.$$

$$\Rightarrow \langle \boldsymbol{\theta}, \mathbf{x} \rangle - f(\mathbf{x}) \geq \langle \boldsymbol{\theta}, \mathbf{y} \rangle - f(\mathbf{y}), \ \forall \mathbf{y}.$$

- (2) $\Leftrightarrow$ (3): by definition of $f^*(\boldsymbol{\theta})$.
- If f is closed, then $f^{**} = f$ so (3) is equivalent to $f^{**}(\mathbf{x}) + f^*(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{x} \rangle$.

Example 1

- $f(x) = \exp(x)$.

Example 1

- $f(x) = \exp(x)$.
- Hence, $f^* = \max_x (x\theta - \exp(x))$.
- Solving the optimization, we have

Example 1

- $f(x) = \exp(x)$.
- Hence, $f^* = \max_x (x\theta - \exp(x))$.
- Solving the optimization, we have
  - $x^* = \ln\theta$, if $\theta > 0$.
  - $x^* = -\infty$ if $\theta = 0$.
  - $x^* = -\infty$, if $\theta < 0$

  Thus,

$$f^*(\theta) = \begin{cases} \theta\ln\theta - \theta & \text{if } \theta > 0 \\ 0 & \text{if } \theta = 0 \\ +\infty & \text{if } \theta < 0 \end{cases}.$$

Example 2: Conjugate of Inner Product

- Let $f(\mathbf{x}) = \langle \mathbf{z}, \mathbf{x} \rangle$, where $\mathbf{z} \in \mathbb{R}^d$, $\mathbf{z} \neq \mathbf{0}$.

Example 2: Conjugate of Inner Product

- Let $f(\mathbf{x}) = \langle \mathbf{z}, \mathbf{x} \rangle$, where $\mathbf{z} \in \mathbb{R}^d$, $\mathbf{z} \neq \mathbf{0}$.
- Then,

$$f^*(\boldsymbol{\theta}) = \sup_{\mathbf{x} \in \mathbb{R}^d} \langle \boldsymbol{\theta} - \mathbf{z}, \mathbf{x} \rangle = \begin{cases} 0 & \text{if } \boldsymbol{\theta} = \mathbf{z} \\ +\infty & \text{otherwise} \end{cases} .$$

Example 3: Conjugate of Hinge Loss

- Let $f(\mathbf{x}) = \max(1 - \langle \mathbf{z}, \mathbf{x} \rangle, 0)$, where $\mathbf{z} \in \mathbb{R}^d$.

Example 3: Conjugate of Hinge Loss

- Let $f(\mathbf{x}) = \max(1 - \langle \mathbf{z}, \mathbf{x} \rangle, 0)$, where $\mathbf{z} \in \mathbb{R}^d$.
- If $\theta = \mathbf{h} + \mathbf{v}$ where $\mathbf{v} \neq \mathbf{0}$, $\mathbf{v} \perp \mathbf{z}$,

Example 3: Conjugate of Hinge Loss

- Let $f(\mathbf{x}) = \max(1 - \langle \mathbf{z}, \mathbf{x} \rangle, 0)$, where $\mathbf{z} \in \mathbb{R}^d$.
- If $\boldsymbol{\theta} = \mathbf{h} + \mathbf{v}$ where $\mathbf{v} \neq \mathbf{0}$, $\mathbf{v} \perp \mathbf{z}$, then we can choose $\mathbf{x}$ along with $\mathbf{v}$.
  - $\langle \mathbf{z}, \mathbf{x} \rangle \to -\infty \Rightarrow \ f^*(\boldsymbol{\theta}) \to \infty$.
- Then, assume that $\boldsymbol{\theta} = \alpha \mathbf{z}$, we have

$$f^*(\boldsymbol{\theta}) = \sup_{\mathbf{x} \in \mathbb{R}^d} \alpha \langle \mathbf{z}, \mathbf{x} \rangle - \max(1 - \langle \mathbf{z}, \mathbf{x} \rangle, 0)$$

Example 3: Conjugate of Hinge Loss

- Let $f(\mathbf{x}) = \max(1 - \langle \mathbf{z}, \mathbf{x} \rangle, 0)$, where $\mathbf{z} \in \mathbb{R}^d$.
- If $\boldsymbol{\theta} = \mathbf{h} + \mathbf{v}$ where $\mathbf{v} \neq \mathbf{0}$, $\mathbf{v} \perp \mathbf{z}$, then we can choose $\mathbf{x}$ along with $\mathbf{v}$.
  - $\langle \mathbf{z}, \mathbf{x} \rangle \to -\infty \Rightarrow f^*(\boldsymbol{\theta}) \to \infty$.
- Then, assume that $\boldsymbol{\theta} = \alpha \mathbf{z}$, we have

$$f^*(\boldsymbol{\theta}) = \sup_{\mathbf{x} \in \mathbb{R}^d} \alpha \langle \mathbf{z}, \mathbf{x} \rangle - \max(1 - \langle \mathbf{z}, \mathbf{x} \rangle, 0) = \sup_u \alpha u - \max(1 - u, 0),$$

so

$$f^*(\boldsymbol{\theta}) = \begin{cases} \alpha, & \text{if } \boldsymbol{\theta} = \alpha \mathbf{z},\ \alpha \in [-1, 0], \\ +\infty, & \text{otherwise.} \end{cases}$$

Example 4: Conjugate of Squared Norms

- Let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, where $\| \cdot \|$ is a norm in $\mathbb{R}^d$, with dual norm $\|\|_*$.

Example 4: Conjugate of Squared Norms

- Let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, where $\|\cdot\|$ is a norm in $\mathbb{R}^d$, with dual norm $\|\|_*$.
- First, note that for all $\mathbf{x}$,

Example 4: Conjugate of Squared Norms

- Let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, where $\|\cdot\|$ is a norm in $\mathbb{R}^d$, with dual norm $\|\|_*$.
- First, note that for all $\mathbf{x}$,

$$\langle \boldsymbol{\theta}, \mathbf{x} \rangle - \frac{1}{2}\|\mathbf{x}\|^2 \leq \|\boldsymbol{\theta}\|_* \|\mathbf{x}\| - \frac{1}{2}\|\mathbf{x}\|^2$$

Example 4: Conjugate of Squared Norms

- Let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, where $\|\cdot\|$ is a norm in $\mathbb{R}^d$, with dual norm $\|\|_*$.
- First, note that for all $\mathbf{x}$,

$$\langle \boldsymbol{\theta}, \mathbf{x} \rangle - \frac{1}{2}\|\mathbf{x}\|^2 \leq \|\boldsymbol{\theta}\|_* \|\mathbf{x}\| - \frac{1}{2}\|\mathbf{x}\|^2$$

- The max value of the right-hand side above (w.r.t. $\|\mathbf{x}\|$):

Example 4: Conjugate of Squared Norms

- Let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, where $\|\cdot\|$ is a norm in $\mathbb{R}^d$, with dual norm $\|\|_*$.
- First, note that for all $\mathbf{x}$,

$$\langle \boldsymbol{\theta}, \mathbf{x} \rangle - \frac{1}{2}\|\mathbf{x}\|^2 \leq \|\boldsymbol{\theta}\|_* \|\mathbf{x}\| - \frac{1}{2}\|\mathbf{x}\|^2$$

- The max value of the right-hand side above (w.r.t. $\|\mathbf{x}\|$): $\frac{1}{2}\|\boldsymbol{\theta}\|_*^2$. So for all $\mathbf{x}$, we have

Example 4: Conjugate of Squared Norms

- Let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, where $\|\cdot\|$ is a norm in $\mathbb{R}^d$, with dual norm $\|\|_*$.
- First, note that for all $\mathbf{x}$,

$$\langle \boldsymbol{\theta}, \mathbf{x} \rangle - \frac{1}{2}\|\mathbf{x}\|^2 \leq \|\boldsymbol{\theta}\|_* \|\mathbf{x}\| - \frac{1}{2}\|\mathbf{x}\|^2$$

- The max value of the right-hand side above (w.r.t. $\|\mathbf{x}\|$): $\frac{1}{2}\|\boldsymbol{\theta}\|_*^2$. So for all $\mathbf{x}$, we have

$$\langle \boldsymbol{\theta}, \mathbf{x} \rangle - \frac{1}{2}\|\mathbf{x}\|^2 \leq \frac{1}{2}\|\boldsymbol{\theta}\|_*^2.$$

  Hence, $f^*(\boldsymbol{\theta}) \leq \frac{1}{2}\|\boldsymbol{\theta}\|_*^2$.

Example 4: Conjugate of Squared Norms

- Let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, where $\|\cdot\|$ is a norm in $\mathbb{R}^d$, with dual norm $\|\|_*$.
- First, note that for all $\mathbf{x}$,

$$\langle \boldsymbol{\theta}, \mathbf{x} \rangle - \frac{1}{2}\|\mathbf{x}\|^2 \leq \|\boldsymbol{\theta}\|_* \|\mathbf{x}\| - \frac{1}{2}\|\mathbf{x}\|^2$$

- The max value of the right-hand side above (w.r.t. $\|\mathbf{x}\|$): $\frac{1}{2}\|\boldsymbol{\theta}\|_*^2$. So for all $\mathbf{x}$, we have

$$\langle \boldsymbol{\theta}, \mathbf{x} \rangle - \frac{1}{2}\|\mathbf{x}\|^2 \leq \frac{1}{2}\|\boldsymbol{\theta}\|_*^2.$$

  Hence, $f^*(\boldsymbol{\theta}) \leq \frac{1}{2}\|\boldsymbol{\theta}\|_*^2$.

- The other side: let $\mathbf{x}$ be any vector with $\langle \boldsymbol{\theta}, \mathbf{x} \rangle = \|\boldsymbol{\theta}\|_* \|\mathbf{x}\|$, scaled with $\|\mathbf{x}\| = \|\boldsymbol{\theta}\|_*$

Example 4: Conjugate of Squared Norms

- Let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, where $\|\cdot\|$ is a norm in $\mathbb{R}^d$, with dual norm $\|\|_*$.
- First, note that for all $\mathbf{x}$,
$$\langle \boldsymbol{\theta}, \mathbf{x}\rangle - \frac{1}{2}\|\mathbf{x}\|^2 \le \|\boldsymbol{\theta}\|_*\|\mathbf{x}\| - \frac{1}{2}\|\mathbf{x}\|^2$$
- The max value of the right-hand side above (w.r.t. $\|\mathbf{x}\|$): $\frac{1}{2}\|\boldsymbol{\theta}\|_*^2$. So for all $\mathbf{x}$, we have
$$\langle \boldsymbol{\theta}, \mathbf{x}\rangle - \frac{1}{2}\|\mathbf{x}\|^2 \le \frac{1}{2}\|\boldsymbol{\theta}\|_*^2.$$
Hence, $f^*(\boldsymbol{\theta}) \le \frac{1}{2}\|\boldsymbol{\theta}\|_*^2$.
- The other side: let $\mathbf{x}$ be any vector with $\langle \boldsymbol{\theta}, \mathbf{x}\rangle = \|\boldsymbol{\theta}\|_*\|\mathbf{x}\|$, scaled with $\|\mathbf{x}\| = \|\boldsymbol{\theta}\|_* \Rightarrow$ we have $f^*(\boldsymbol{\theta}) \ge \frac{1}{2}\|\boldsymbol{\theta}\|_*^2$.

Example 5: Conjugate of Affine Functions

Lemma

Let f be a function and f^* be its Fenchel conjugate. For $a > 0$ and $b \in \mathbb{R}$, the Fenchel conjugate of $g(\mathbf{x}) = af(\mathbf{x}) + b$ is

$$g^*(\boldsymbol{\theta}) = af^*(\boldsymbol{\theta}/a) - b.$$

- By definition,

$$\begin{aligned} g^*(\boldsymbol{\theta}) &= \sup_{\mathbf{x}\in\mathbb{R}^d} \left(\langle \boldsymbol{\theta}, \mathbf{x}\rangle - af(\mathbf{x}) - b\right) \\ &= -b + a\left(\sup_{\mathbf{x}\in\mathbb{R}^d}\left(\left\langle \frac{\boldsymbol{\theta}}{a}, \mathbf{x}\right\rangle - f(\mathbf{x})\right)\right) \\ &= -b + af^*(\boldsymbol{\theta}/a). \end{aligned}$$

Example 6: Order Reverse

Lemma

Let f_1, f_2 be two functions such that $f_1(\mathbf{x}) \leq f_2(\mathbf{x})$ for all $\mathbf{x}$. Then, $f_1^*(\boldsymbol{\theta}) \geq f_2^*(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$.

Example 6: Order Reverse

Lemma

Let f_1, f_2 be two functions such that $f_1(\mathbf{x}) \leq f_2(\mathbf{x})$ for all $\mathbf{x}$. Then, $f_1^*(\boldsymbol{\theta}) \geq f_2^*(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$.

- By definition,

$$f_1^*(\boldsymbol{\theta}) = \sup_{\mathbf{x}} \left(\langle \boldsymbol{\theta}, \mathbf{x} \rangle - f_1(\mathbf{x}) \right)$$

Example 6: Order Reverse

Lemma

Let f_1, f_2 be two functions such that $f_1(\mathbf{x}) \leq f_2(\mathbf{x})$ for all $\mathbf{x}$. Then, $f_1^*(\boldsymbol{\theta}) \geq f_2^*(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$.

- By definition,

$$f_1^*(\boldsymbol{\theta}) = \sup_{\mathbf{x}} \left(\langle \boldsymbol{\theta}, \mathbf{x} \rangle - f_1(\mathbf{x}) \right) \geq \sup_{\mathbf{x}} \left(\langle \boldsymbol{\theta}, \mathbf{x} \rangle - f_2(\mathbf{x}) \right) = f_2^*(\boldsymbol{\theta}).$$

Discussions