Online Learning

— Online (Sub-)Gradient Descent with Strong Convexity

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

Strongly Convex Function

Strongly Convex Function

Let $\mu \geq 0$. A function $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$ is μ-strongly convex over a convex set $V \subseteq \mathrm{dom}(\partial f)$ w.r.t. $\|\cdot\|$ if

$$\forall \mathbf{x}, \mathbf{y} \in V, \mathbf{g} \in \partial f(\mathbf{x}),\; f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle + \frac{u}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

Strongly Convex Function

Strongly Convex Function

Let $\mu \geq 0$. A function $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$ is μ-strongly convex over a convex set $V \subseteq \mathrm{dom}(\partial f)$ w.r.t. $\|\cdot\|$ if

$$\forall \mathbf{x}, \mathbf{y} \in V, \mathbf{g} \in \partial f(\mathbf{x}),\ f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle + \frac{u}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

- Taylor series up to the quadratic term.

Strongly Convex Function

Strongly Convex Function

Let $\mu \geq 0$. A function $f : \mathbb{R}^d \mapsto (-\infty, +\infty]$ is μ-strongly convex over a convex set $V \subseteq \text{dom}(\partial f)$ w.r.t. $\|\cdot\|$ if

$$\forall \mathbf{x}, \mathbf{y} \in V, \mathbf{g} \in \partial f(\mathbf{x}),\ f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle + \frac{u}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

- Taylor series up to the quadratic term.
- For twice differentiable functions, we have the following theorem, which is useful.

Strongly Convex Function

Theorem [Shalev-Shwartz, 2007]

Let $V \subseteq \mathbb{R}^d$ be a convex set and $f : V \mapsto \mathbb{R}$ be a twice differentiable function. Then f is μ-strongly convex in V w.r.t. $\|\|$ if for all $\mathbf{x}, \mathbf{y} \in V$, we have

$$\langle \nabla^2 f(\mathbf{x})\mathbf{y}, \mathbf{y} \rangle \geq \mu \|\mathbf{y}\|^2,$$

where $\nabla^2 f(\mathbf{x})$ is the Hessian matrix of f at $\mathbf{x}$.

- That is, $\nabla^2 f(\mathbf{x}) \succeq \mu I$.
- Further readings: [link].

Strong Convexity is Additive

Theorem

Given two functions f, g which are strongly convex in a non-empty convex set $V \subseteq \text{int dom}(f) \cap \text{int dom}(g)$ w.r.t. $\|\cdot\|$, and

- $f : \mathbb{R}^d \mapsto \mathbb{R}$ is μ_1-strongly convex
- $g : \mathbb{R}^d \mapsto \mathbb{R}$ is μ_2-strongly convex

Then, $f + g$ is $(\mu_1 + \mu_2)$-strongly convex in V w.r.t. $\|\cdot\|$.

An Exericse

Exercise

Show that $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ is 1-strongly convex w.r.t. $\|\cdot\|_2$ in $\mathbf{R}^d$.

An Exericse

Exercise

Show that $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ is 1-strongly convex w.r.t. $\|\cdot\|_2$ in $\mathbf{R}^d$.

- *Hint:* Apply the theorem by Shalev & Shwartz.

Outline

Recall: Online (Sub-)Gradient Descent (GD)

1. **Input:** convex set V, T, $\mathbf{x}_1 \in V$, step size $\{\eta_t\}$.
2. **for** $t \leftarrow 1$ to T **do**:
   1. Play $\mathbf{x}_t$ and observe cost $f_t(\mathbf{x}_t)$.
   2. Update and Project:

$$
\begin{aligned}
\mathbf{y}_{t+1} &= \mathbf{x}_t - \eta_t \mathbf{g}_t, \text{ for } \mathbf{g}_t \in \partial f_t(\mathbf{x}_t) \\
\mathbf{x}_{t+1} &= \Pi_{\mathcal{K}}(\mathbf{y}_{t+1})
\end{aligned}
$$

3. **end for**

Steps for the regret bound (1/5)

- Consider $\|\cdot\| = \|\cdot\|_2$.
- For a fixed $\mathbf{u} \in V$, we have

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{u}\|^2 - \|\mathbf{x}_t - \mathbf{u}\|^2 &\leq \|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{u}\|^2 - \|\mathbf{x}_t - \mathbf{u}\|^2 \\
&= -2\eta_t \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle + \eta_t^2 \|\mathbf{g}_t\|^2 \\
&\leq -2\eta_t (f_t(\mathbf{x}_t) - f_t(\mathbf{u})) + \eta_t^2 \|\mathbf{g}_t\|^2.
\end{aligned}$$

Steps for the regret bound (1/5)

- Consider $\|\cdot\| = \|\cdot\|_2$.
- For a fixed $\mathbf{u} \in V$, we have

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{u}\|^2 - \|\mathbf{x}_t - \mathbf{u}\|^2 &\leq \|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{u}\|^2 - \|\mathbf{x}_t - \mathbf{u}\|^2 \\
&= -2\eta_t \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle + \eta_t^2 \|\mathbf{g}_t\|^2 \\
&\leq -2\eta_t (f_t(\mathbf{x}_t) - f_t(\mathbf{u})) + \eta_t^2 \|\mathbf{g}_t\|^2.
\end{aligned}$$

Hence we derive that

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

Steps for the regret bound (1/5)

- Consider $\|\cdot\| = \|\cdot\|_2$.
- For a fixed $\mathbf{u} \in V$, we have

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{u}\|^2 - \|\mathbf{x}_t - \mathbf{u}\|^2 &\le \|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{u}\|^2 - \|\mathbf{x}_t - \mathbf{u}\|^2 \\ &= -2\eta_t \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle + \eta_t^2 \|\mathbf{g}_t\|^2 \\ &\le -2\eta_t (f_t(\mathbf{x}_t) - f_t(\mathbf{u})) + \eta_t^2 \|\mathbf{g}_t\|^2.\end{aligned}$$

Hence we derive that

$$\begin{aligned}f_t(\mathbf{x}_t) - f_t(\mathbf{u}) &\le \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle \\ &\le \frac{1}{2\eta_t}\|\mathbf{x}_t - \mathbf{u}\|^2 - \frac{1}{2\eta_t}\|\mathbf{x}_{t+1} - \mathbf{u}\|^2 + \frac{\eta_t}{2}\|\mathbf{g}_t\|^2.\end{aligned}$$

Steps for the regret bound (1/5)

- Consider $\|\cdot\| = \|\cdot\|_2$.
- For a fixed $\mathbf{u} \in V$, we have

$$\begin{aligned} \|\mathbf{x}_{t+1} - \mathbf{u}\|^2 - \|\mathbf{x}_t - \mathbf{u}\|^2 &\leq \|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{u}\|^2 - \|\mathbf{x}_t - \mathbf{u}\|^2 \\ &= -2\eta_t \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle + \eta_t^2 \|\mathbf{g}_t\|^2 \\ &\leq -2\eta_t (f_t(\mathbf{x}_t) - f_t(\mathbf{u})) + \eta_t^2 \|\mathbf{g}_t\|^2. \end{aligned}$$

Hence we derive that

$$\begin{aligned} f_t(\mathbf{x}_t) - f_t(\mathbf{u}) &\leq \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle \\ &\leq \frac{1}{2\eta_t}\|\mathbf{x}_t - \mathbf{u}\|^2 - \frac{1}{2\eta_t}\|\mathbf{x}_{t+1} - \mathbf{u}\|^2 + \frac{\eta_t}{2}\|\mathbf{g}_t\|^2. \end{aligned}$$

Steps for the regret bound (2/5)

- Suppose $f_t : \mathbb{R}^d \mapsto \mathbb{R}$ is μ_t-strongly convex w.r.t. $\|\cdot\|_2$ over $V \subseteq \text{int dom}(f_t)$ for $\mu_t > 0$, $\forall t$.
- The strong convexity leads to

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle - \frac{\mu_t}{2}\|\mathbf{x}_t - \mathbf{u}\|^2.$$

Steps for the regret bound (3/5)

- We can set the learning rate adaptively by $\eta_t = 1/(\sum_{i=1}^{t} \mu_i)$.
- So we have

$$
\begin{aligned}
\frac{1}{2\eta_1} - \frac{\mu_1}{2} &= 0 \\
\frac{1}{2\eta_t} - \frac{\mu_t}{2} &= \frac{1}{2\eta_{t-1}}, \text{ for } t \geq 2.
\end{aligned}
$$

Steps for the regret bound (3/5)

- We can set the learning rate adaptively by $\eta_t = 1/(\sum_{i=1}^{t} \mu_i)$.
- So we have

$$
\begin{aligned}
\frac{1}{2\eta_1} - \frac{\mu_1}{2} &= 0 \\
\frac{1}{2\eta_t} - \frac{\mu_t}{2} &= \frac{1}{2\eta_{t-1}}, \text{ for } t \geq 2.
\end{aligned}
$$

⋆ The learning rate is getting smaller with time.

Steps for the regret bound (4/5)

- Summing up the previous regret bound:

$$\sum_{t=1}^{T}(f_t(\mathbf{x}_t) - f_t(\mathbf{u})) \leq \sum_{t=1}^{T}\left(\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u}\rangle - \frac{\mu_t}{2}\|\mathbf{x}_t - \mathbf{u}\|^2\right)$$

Steps for the regret bound (4/5)

- Summing up the previous regret bound:

$$\sum_{t=1}^{T}(f_t(\mathbf{x}_t) - f_t(\mathbf{u})) \leq \sum_{t=1}^{T}\left(\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u}\rangle - \frac{\mu_t}{2}\|\mathbf{x}_t - \mathbf{u}\|^2\right)$$

$$\leq \sum_{t=1}^{T}\left(\frac{1}{2\eta_t}\|\mathbf{x}_t - \mathbf{u}\|^2 - \frac{1}{2\eta_t}\|\mathbf{x}_{t+1} - \mathbf{u}\|^2 + \frac{\eta_t}{2}\|\mathbf{g}_t\|^2 - \frac{\mu_t}{2}\|\mathbf{x}_t - \mathbf{u}\|^2\right)$$

Steps for the regret bound (4/5)

- Summing up the previous regret bound:

$$\sum_{t=1}^{T}(f_t(\mathbf{x}_t) - f_t(\mathbf{u})) \leq \sum_{t=1}^{T}\left(\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u}\rangle - \frac{\mu_t}{2}\|\mathbf{x}_t - \mathbf{u}\|^2\right)$$

$$\leq \sum_{t=1}^{T}\left(\frac{1}{2\eta_t}\|\mathbf{x}_t - \mathbf{u}\|^2 - \frac{1}{2\eta_t}\|\mathbf{x}_{t+1} - \mathbf{u}\|^2 + \frac{\eta_t}{2}\|\mathbf{g}_t\|^2 - \frac{\mu_t}{2}\|\mathbf{x}_t - \mathbf{u}\|^2\right)$$

$$= -\frac{1}{2\eta_1}\|\mathbf{x}_2 - \mathbf{u}\|^2 + \sum_{t=2}^{T}\left(\frac{1}{2\eta_{t-1}}\|\mathbf{x}_t - \mathbf{u}\|^2 - \frac{1}{2\eta_t}\|\mathbf{x}_{t+1} - \mathbf{u}\|^2\right)$$

$$+ \sum_{t=1}^{T}\frac{\eta_t}{2}\|\mathbf{g}_t\|^2$$

Steps for the regret bound (4/5)

- Summing up the previous regret bound:

$$
\begin{aligned}
&\sum_{t=1}^{T}(f_t(\mathbf{x}_t) - f_t(\mathbf{u})) \le \sum_{t=1}^{T}\left(\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u}\rangle - \frac{\mu_t}{2}\|\mathbf{x}_t - \mathbf{u}\|^2\right) \\
\le \quad & \sum_{t=1}^{T}\left(\frac{1}{2\eta_t}\|\mathbf{x}_t - \mathbf{u}\|^2 - \frac{1}{2\eta_t}\|\mathbf{x}_{t+1} - \mathbf{u}\|^2 + \frac{\eta_t}{2}\|\mathbf{g}_t\|^2 - \frac{\mu_t}{2}\|\mathbf{x}_t - \mathbf{u}\|^2\right) \\
= \quad & -\frac{1}{2\eta_1}\|\mathbf{x}_2 - \mathbf{u}\|^2 + \sum_{t=2}^{T}\left(\frac{1}{2\eta_{t-1}}\|\mathbf{x}_t - \mathbf{u}\|^2 - \frac{1}{2\eta_t}\|\mathbf{x}_{t+1} - \mathbf{u}\|^2\right) \\
& + \sum_{t=1}^{T}\frac{\eta_t}{2}\|\mathbf{g}_t\|^2 \\
\le \quad & \sum_{t=1}^{T}\frac{\eta_t}{2}\|\mathbf{g}_t\|^2.
\end{aligned}
$$

Steps for the regret bound (4/5)

- Further assumptions:
    - $\mu_t = \mu > 0$ for all t.
    - f_t is L-Lipschitz w.r.t. $\|\cdot\| = \|\cdot\|_2$ for all t.
    - Set the learning rate adaptively by $\eta_t = 1/(\sum_{i=1}^{t} \mu_i)$.

Steps for the regret bound (4/5)

- Further assumptions:
  - $\mu_t = \mu > 0$ for all t.
  - f_t is L-Lipschitz w.r.t. $\|\cdot\| = \|\cdot\|_2$ for all t.
  - Set the learning rate adaptively by $\eta_t = 1/(\sum_{i=1}^{t} \mu_i)$.
- Then we have

$$\begin{aligned}\sum_{t=1}^{T}(f_t(\mathbf{x}_t) - f_t(\mathbf{u})) &\leq \sum_{t=1}^{T} \frac{\eta_t}{2}\|\mathbf{g}_t\|^2 \\ &= \sum_{t=1}^{T} \frac{1}{2\sum_{i=1}^{t}\mu_i}\|\mathbf{g}_t\|^2 \\ &\leq \frac{L^2}{2\mu}(1 + \ln T).\end{aligned}$$

Discussions