Online Learning

— Online Gradient Descent & Subgradients

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

Why so complicated?

- How about just *following the one with best performance*?

Why so complicated?

- How about just *following the one with best performance*?
  - Follow The Leader (FTL) Algorithm.

Why so complicated?

- How about just *following the one with best performance*?
  - Follow The Leader (FTL) Algorithm.

- First, we assume to make no assumptions on $\mathcal{K}$ and $\{f_t : L \mapsto \mathbb{R}\}$.
- At time t, we are given previous cost functions $f_1, \ldots, f_{t-1}$, and then give the solution

$$\mathbf{x}_t := \arg\min_{\mathbf{x} \in \mathcal{K}} \sum_{k=1}^{t-1} f_k(\mathbf{x}).$$

Why so complicated?

- How about just *following the one with best performance*?
  - Follow The Leader (FTL) Algorithm.

- First, we assume to make no assumptions on $\mathcal{K}$ and $\{f_t : L \mapsto \mathbb{R}\}$.
- At time t, we are given previous cost functions $f_1, \ldots, f_{t-1}$, and then give the solution

$$\mathbf{x}_t := \arg\min_{\mathbf{x}\in\mathcal{K}} \sum_{k=1}^{t-1} f_k(\mathbf{x}).$$

That is, the best solution for the previous $t-1$ steps.

Why so complicated?

- How about just *following the one with best performance*?
  - Follow The Leader (FTL) Algorithm.

- First, we assume to make no assumptions on $\mathcal{K}$ and $\{f_t : L \mapsto \mathbb{R}\}$.
- At time t, we are given previous cost functions $f_1, \ldots, f_{t-1}$, and then give the solution

$$\mathbf{x}_t := \arg\min_{\mathbf{x}\in\mathcal{K}} \sum_{k=1}^{t-1} f_k(\mathbf{x}).$$

  That is, the best solution for the previous $t-1$ steps.

- It seems reasonable and makes sense, doesn't it?

FTL leads to "overfitting"

t:	1
$\mathbf{x}_t$:	$(0.5, 0.5)$
ℓ_t:	$(0, 0.5)$
$f_t(\mathbf{x}_t)$:	0.25
$\arg\min_{\mathbf{x}} \sum_{k=1}^{t} f_k(\mathbf{x})$:	$(1, 0)$

FTL leads to "overfitting"

t:	1	2
$\mathbf{x}_t$:	$(0.5, 0.5)$	$(1, 0)$
ℓ_t:	$(0, 0.5)$	$(1, 0)$
$f_t(\mathbf{x}_t)$:	0.25	1
$\arg\min_{\mathbf{x}} \sum_{k=1}^{t} f_k(\mathbf{x})$:	$(1, 0)$	$(0, 1)$

FTL leads to “overfitting”

t:	1	2	3
$\mathbf{x}_t$:	$(0.5, 0.5)$	$(1, 0)$	$(0, 1)$
ℓ_t:	$(0, 0.5)$	$(1, 0)$	$(0, 1)$
$f_t(\mathbf{x}_t)$:	0.25	1	1
$\arg\min_{\mathbf{x}} \sum_{k=1}^{t} f_k(\mathbf{x})$:	$(1, 0)$	$(0, 1)$	$(1, 0)$

FTL leads to "overfitting"

t:	1	2	3	4
$\mathbf{x}_t$:	$(0.5, 0.5)$	$(1, 0)$	$(0, 1)$	$(1, 0)$
ℓ_t:	$(0, 0.5)$	$(1, 0)$	$(0, 1)$	$(1, 0)$
$f_t(\mathbf{x}_t)$:	0.25	1	1	1
$\arg\min_{\mathbf{x}} \sum_{k=1}^{t} f_k(\mathbf{x})$:	$(1, 0)$	$(0, 1)$	$(1, 0)$	$(0, 1)$

FTL leads to “overfitting”

t:	1	2	3	4	5
$\mathbf{x}_t$:	$(0.5, 0.5)$	$(1, 0)$	$(0, 1)$	$(1, 0)$	$(0, 1)$
ℓ_t:	$(0, 0.5)$	$(1, 0)$	$(0, 1)$	$(1, 0)$	$(0, 1)$
$f_t(\mathbf{x}_t)$:	0.25	1	1	1	1
$\arg\min_{\mathbf{x}} \sum_{k=1}^{t} f_k(\mathbf{x})$:	$(1, 0)$	$(0, 1)$	$(1, 0)$	$(0, 1)$	$(1, 0)$

FTL leads to "overfitting"

t:	1	2	3	4	5	...
$\mathbf{x}_t$:	$(0.5, 0.5)$	$(1, 0)$	$(0, 1)$	$(1, 0)$	$(0, 1)$	...
ℓ_t:	$(0, 0.5)$	$(1, 0)$	$(0, 1)$	$(1, 0)$	$(0, 1)$	...
$f_t(\mathbf{x}_t)$:	0.25	1	1	1	1	...
$\arg\min_{\mathbf{x}} \sum_{k=1}^{t} f_k(\mathbf{x})$:	$(1, 0)$	$(0, 1)$	$(1, 0)$	$(0, 1)$	$(1, 0)$	...

FTL leads to "overfitting"

t:	1	2	3	4	5	...
$\mathbf{x}_t$:	$(0.5, 0.5)$	$(1, 0)$	$(0, 1)$	$(1, 0)$	$(0, 1)$	...
ℓ_t:	$(0, 0.5)$	$(1, 0)$	$(0, 1)$	$(1, 0)$	$(0, 1)$	...
$f_t(\mathbf{x}_t)$:	0.25	1	1	1	1	...
$\arg\min_{\mathbf{x}} \sum_{k=1}^{t} f_k(\mathbf{x})$:	$(1, 0)$	$(0, 1)$	$(1, 0)$	$(0, 1)$	$(1, 0)$	...

optimum loss: $\approx T/2$.

FTL's loss: $\approx T$.

regret: $\approx T/2$ (linear).

Remark

- Note that the first example of no-regret analysis in this course uses a special kind of loss function.
  - Squared difference: $\|\mathbf{x}_t - \mathbf{y}_t\|_2^2$.

Outline

Online Gradient Descent (GD)

1. **Input:** convex set $\mathcal{K}$, T, $\mathbf{x}_1 \in \mathcal{K}$, learning rate $\{\eta_t\}$.
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:

$$\begin{aligned} \mathbf{y}_{t+1} &= \mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \\ \mathbf{x}_{t+1} &= \Pi_{\mathcal{K}}(\mathbf{y}_{t+1}) \end{aligned}$$

3. **end for**

GD for online convex optimization is of no-regret

Theorem A

Online gradient descent with learning rate $\{\eta_t = \frac{D}{G\sqrt{t}}, t \in [T]\}$ guarantees the following for all $T \geq 1$:

$$\text{regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}^* \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}^*) \leq \frac{3}{2} GD\sqrt{T}.$$

- D: the diameter of $\mathcal{K}$.
- Assume that $\nabla f_t(\mathbf{x}) \leq G$ for each $\mathbf{x} \in \mathcal{K}$.

Proof of Theorem A (1/4)

- Let $\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x})$.
- Since f_t is convex, we have

$$f_t(\mathbf{x}_t) - f_t(\mathbf{x}^*) \leq (\nabla f_t(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*).$$

- By the updating rule for $\mathbf{x}_{t+1}$ and the Pythagorean theorem, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 = \|\Pi_{\mathcal{K}}(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t)) - \mathbf{x}^*\|^2 \leq \|\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) - \mathbf{x}^*\|^2.$$

Proof of Theorem A (2/4)

- Hence

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \eta_t^2 \|\nabla f_t(\mathbf{x}_t)\|^2 - 2\eta_t (\nabla f_t(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*)$$

$$2(\nabla f_t(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*) \leq \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2}{\eta_t} + \eta_t G^2.$$

- Summing above inequality from $t = 1$ to T and setting $\eta_t = \frac{D}{G\sqrt{t}}$ and $\frac{1}{\eta_0} := 0$ we have :

Proof of Theorem A (3/4)

$$
\begin{aligned}
2\left(\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}^*)\right) &\leq 2\sum_{t=1}^{T} \nabla f_t(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*) \\
&\leq \sum_{t=1}^{T} \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2}{\eta_t} + G^2 \sum_{t=1}^{T} \eta_t \\
&\leq \sum_{t=1}^{T} \|\mathbf{x}_t - \mathbf{x}^*\|^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + G^2 \sum_{t=1}^{T} \eta_t \\
&\leq D^2 \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + G^2 \sum_{t=1}^{T} \eta_t \\
&\leq D^2 \frac{1}{\eta_T} + G^2 \sum_{t=1}^{T} \eta_t \\
&\leq 3DG\sqrt{T}.
\end{aligned}
$$

Proof of Theorem A (4/4)

Note that we can also deduce in this way

$$\sum_{t=1}^{T} \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2}{\eta_t} + G^2 \sum_{t=1}^{T} \eta_t$$

Proof of Theorem A (4/4)

Note that we can also deduce in this way

$$
\begin{aligned}
&\sum_{t=1}^{T} \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2}{\eta_t} + G^2 \sum_{t=1}^{T} \eta_t \\
= \quad & \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{\eta_1} + \|\mathbf{x}_2 - \mathbf{x}^*\| \left(\frac{1}{\eta_2} - \frac{1}{\eta_1} \right) + \|\mathbf{x}_3 - \mathbf{x}^*\| \left(\frac{1}{\eta_3} - \frac{1}{\eta_2} \right) + \cdots
\end{aligned}
$$

Proof of Theorem A (4/4)

Note that we can also deduce in this way

$$
\begin{aligned}
& \sum_{t=1}^{T} \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2}{\eta_t} + G^2 \sum_{t=1}^{T} \eta_t \\
= \quad & \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{\eta_1} + \|\mathbf{x}_2 - \mathbf{x}^*\| \left(\frac{1}{\eta_2} - \frac{1}{\eta_1} \right) + \|\mathbf{x}_3 - \mathbf{x}^*\| \left(\frac{1}{\eta_3} - \frac{1}{\eta_2} \right) + \cdots \\
& + \|\mathbf{x}_T - \mathbf{x}^*\| \left(\frac{1}{\eta_T} - \frac{1}{\eta_{T-1}} \right) - \frac{\|\mathbf{x}_T - \mathbf{x}^*\|}{\eta_T} + G^2 \sum_{t=1}^{T} \eta_t
\end{aligned}
$$

Proof of Theorem A (4/4)

Note that we can also deduce in this way

$$
\begin{aligned}
&\sum_{t=1}^{T} \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2}{\eta_t} + G^2 \sum_{t=1}^{T} \eta_t \\
= \quad & \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{\eta_1} + \|\mathbf{x}_2 - \mathbf{x}^*\| \left(\frac{1}{\eta_2} - \frac{1}{\eta_1}\right) + \|\mathbf{x}_3 - \mathbf{x}^*\| \left(\frac{1}{\eta_3} - \frac{1}{\eta_2}\right) + \cdots \\
& + \|\mathbf{x}_T - \mathbf{x}^*\| \left(\frac{1}{\eta_T} - \frac{1}{\eta_{T-1}}\right) - \frac{\|\mathbf{x}_T - \mathbf{x}^*\|}{\eta_T} + G^2 \sum_{t=1}^{T} \eta_t \\
\leq \quad & \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{\eta_1} + D^2 \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + G^2 \sum_{t=1}^{T} \eta_t
\end{aligned}
$$

Proof of Theorem A (4/4)

Note that we can also deduce in this way

$$
\begin{aligned}
&\sum_{t=1}^{T} \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2}{\eta_t} + G^2 \sum_{t=1}^{T} \eta_t \\
= \quad & \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{\eta_1} + \|\mathbf{x}_2 - \mathbf{x}^*\| \left(\frac{1}{\eta_2} - \frac{1}{\eta_1} \right) + \|\mathbf{x}_3 - \mathbf{x}^*\| \left(\frac{1}{\eta_3} - \frac{1}{\eta_2} \right) + \cdots \\
& + \|\mathbf{x}_T - \mathbf{x}^*\| \left(\frac{1}{\eta_T} - \frac{1}{\eta_{T-1}} \right) - \frac{\|\mathbf{x}_T - \mathbf{x}^*\|}{\eta_T} + G^2 \sum_{t=1}^{T} \eta_t \\
\leq \quad & \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{\eta_1} + D^2 \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) + G^2 \sum_{t=1}^{T} \eta_t \\
\leq \quad & \frac{D^2}{\eta_1} + D^2 \left(\frac{1}{\eta_T} - \frac{1}{\eta_1} \right) + G^2 \sum_{t=1}^{T} \eta_t
\end{aligned}
$$

Proof of Theorem A (4/4)

Note that we can also deduce in this way

$$
\begin{aligned}
& \sum_{t=1}^{T} \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2}{\eta_t} + G^2 \sum_{t=1}^{T} \eta_t \\
= \quad & \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{\eta_1} + \|\mathbf{x}_2 - \mathbf{x}^*\| \left(\frac{1}{\eta_2} - \frac{1}{\eta_1} \right) + \|\mathbf{x}_3 - \mathbf{x}^*\| \left(\frac{1}{\eta_3} - \frac{1}{\eta_2} \right) + \cdots \\
& + \|\mathbf{x}_T - \mathbf{x}^*\| \left(\frac{1}{\eta_T} - \frac{1}{\eta_{T-1}} \right) - \frac{\|\mathbf{x}_T - \mathbf{x}^*\|}{\eta_T} + G^2 \sum_{t=1}^{T} \eta_t \\
\leq \quad & \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{\eta_1} + D^2 \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) + G^2 \sum_{t=1}^{T} \eta_t \\
\leq \quad & \frac{D^2}{\eta_1} + D^2 \left(\frac{1}{\eta_T} - \frac{1}{\eta_1} \right) + G^2 \sum_{t=1}^{T} \eta_t \\
= \quad & \frac{D^2}{\eta_T} + G^2 \sum_{t=1}^{T} \frac{D}{G\sqrt{t}} \leq
\end{aligned}
$$

Proof of Theorem A (4/4)

Note that we can also deduce in this way

$$
\begin{aligned}
& \sum_{t=1}^{T} \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2}{\eta_t} + G^2 \sum_{t=1}^{T} \eta_t \\
= \quad & \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{\eta_1} + \|\mathbf{x}_2 - \mathbf{x}^*\| \left(\frac{1}{\eta_2} - \frac{1}{\eta_1} \right) + \|\mathbf{x}_3 - \mathbf{x}^*\| \left(\frac{1}{\eta_3} - \frac{1}{\eta_2} \right) + \cdots \\
& + \|\mathbf{x}_T - \mathbf{x}^*\| \left(\frac{1}{\eta_T} - \frac{1}{\eta_{T-1}} \right) - \frac{\|\mathbf{x}_T - \mathbf{x}^*\|}{\eta_T} + G^2 \sum_{t=1}^{T} \eta_t \\
\leq \quad & \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{\eta_1} + D^2 \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) + G^2 \sum_{t=1}^{T} \eta_t \\
\leq \quad & \frac{D^2}{\eta_1} + D^2 \left(\frac{1}{\eta_T} - \frac{1}{\eta_1} \right) + G^2 \sum_{t=1}^{T} \eta_t \\
= \quad & \frac{D^2}{\eta_T} + G^2 \sum_{t=1}^{T} \frac{D}{G\sqrt{t}} \leq \frac{D^2}{D/(G\sqrt{T})} + DG(2\sqrt{T} - 1)
\end{aligned}
$$

Proof of Theorem A (4/4)

Note that we can also deduce in this way

$$
\begin{aligned}
& \sum_{t=1}^{T} \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2}{\eta_t} + G^2 \sum_{t=1}^{T} \eta_t \\
= \quad & \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{\eta_1} + \|\mathbf{x}_2 - \mathbf{x}^*\| \left(\frac{1}{\eta_2} - \frac{1}{\eta_1} \right) + \|\mathbf{x}_3 - \mathbf{x}^*\| \left(\frac{1}{\eta_3} - \frac{1}{\eta_2} \right) + \cdots \\
& + \|\mathbf{x}_T - \mathbf{x}^*\| \left(\frac{1}{\eta_T} - \frac{1}{\eta_{T-1}} \right) - \frac{\|\mathbf{x}_T - \mathbf{x}^*\|}{\eta_T} + G^2 \sum_{t=1}^{T} \eta_t \\
\leq \quad & \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{\eta_1} + D^2 \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) + G^2 \sum_{t=1}^{T} \eta_t \\
\leq \quad & \frac{D^2}{\eta_1} + D^2 \left(\frac{1}{\eta_T} - \frac{1}{\eta_1} \right) + G^2 \sum_{t=1}^{T} \eta_t \\
= \quad & \frac{D^2}{\eta_T} + G^2 \sum_{t=1}^{T} \frac{D}{G\sqrt{t}} \leq \frac{D^2}{D/(G\sqrt{T})} + DG(2\sqrt{T} - 1) \leq 3DG\sqrt{T}.
\end{aligned}
$$

The Lower Bound (for OLO)

Theorem B

Let $\mathcal{K} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_\infty \leq r\}$ be a convex subset of $\mathbb{R}^d$. Let A be any algorithm for Online Linear Optimization on $\mathcal{K}$. Then for any $T \geq 1$, there exists a sequence of vectors $\mathbf{g}_1, \ldots, \mathbf{g}_T$ with $\|\mathbf{g}_t\|_2 \leq L$ and $\mathbf{u} \in \mathcal{K}$ such that the regret of A satisfies

$$\mathrm{regret}_T(\mathbf{u}) = \sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{x}_t \rangle - \sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{u} \rangle \geq \frac{\sqrt{2} L D \sqrt{T}}{4}.$$

- The diameter D of $\mathcal{K}$ is at most $\sqrt{\sum_{i=1}^{d} (2r)^2} \leq 2r\sqrt{d}$.
- $\|\mathbf{x}\|_\infty \leq r \Leftrightarrow |\mathbf{x}(i)| \leq r$ for each $i \in [n]$.

Proof of Theorem B (1/2)

- The approach:

  For any random variable $\mathbf{z}$ with domain $\mathcal{V}$ and any function f,

$$\sup_{\mathbf{x} \in V} f(\mathbf{x}) \geq E[f(\mathbf{z})].$$

Proof of Theorem B (1/2)

- The approach:

  For any random variable $\mathbf{z}$ with domain $\mathcal{V}$ and any function f,

$$\sup_{\mathbf{x}\in V} f(\mathbf{x}) \geq E[f(\mathbf{z})].$$

- $\text{regret}_T = \max_{\mathbf{u}\in\mathcal{K}} \text{regret}_T(\mathbf{u})$.

Proof of Theorem B (1/2)

- The approach:

  For any random variable $\mathbf{z}$ with domain $\mathcal{V}$ and any function f,

$$\sup_{\mathbf{x} \in V} f(\mathbf{x}) \geq E[f(\mathbf{z})].$$

- $\text{regret}_T = \max_{\mathbf{u} \in \mathcal{K}} \text{regret}_T(\mathbf{u})$.
- Let $\mathbf{v}, \mathbf{w} \in \mathcal{K}$ such that $\|\mathbf{v} - \mathbf{w}\| = D$.

Proof of Theorem B (1/2)

- The approach:

  For any random variable $\mathbf{z}$ with domain $\mathcal{V}$ and any function f,

$$\sup_{\mathbf{x}\in V} f(\mathbf{x}) \geq E[f(\mathbf{z})].$$

- $\text{regret}_T = \max_{\mathbf{u}\in\mathcal{K}} \text{regret}_T(\mathbf{u})$.
- Let $\mathbf{v}, \mathbf{w} \in \mathcal{K}$ such that $\|\mathbf{v} - \mathbf{w}\| = D$.
- Let $\mathbf{z} := \frac{\mathbf{v}-\mathbf{w}}{\|\mathbf{v}-\mathbf{w}\|}$

Proof of Theorem B (1/2)

- The approach:

  For any random variable $\mathbf{z}$ with domain $\mathcal{V}$ and any function f,

$$\sup_{\mathbf{x} \in V} f(\mathbf{x}) \geq E[f(\mathbf{z})].$$

- $\text{regret}_T = \max_{\mathbf{u} \in \mathcal{K}} \text{regret}_T(\mathbf{u})$.
- Let $\mathbf{v}, \mathbf{w} \in \mathcal{K}$ such that $\|\mathbf{v} - \mathbf{w}\| = D$.
- Let $\mathbf{z} := \frac{\mathbf{v}-\mathbf{w}}{\|\mathbf{v}-\mathbf{w}\|} \Rightarrow \langle \mathbf{z}, \mathbf{v} - \mathbf{w} \rangle = D$.
- Let $\epsilon_1, \epsilon_2, \ldots, \epsilon_T$ be i.i.d. random variables such that $\Pr[\epsilon_t = 1] = \Pr[\epsilon_t = -1] = 1/2$ for each t.

Proof of Theorem B (1/2)

- The approach:

  For any random variable $\mathbf{z}$ with domain $\mathcal{V}$ and any function f,

$$\sup_{\mathbf{x} \in V} f(\mathbf{x}) \geq E[f(\mathbf{z})].$$

- $\text{regret}_T = \max_{\mathbf{u} \in \mathcal{K}} \text{regret}_T(\mathbf{u})$.
- Let $\mathbf{v}, \mathbf{w} \in \mathcal{K}$ such that $\|\mathbf{v} - \mathbf{w}\| = D$.
- Let $\mathbf{z} := \frac{\mathbf{v}-\mathbf{w}}{\|\mathbf{v}-\mathbf{w}\|} \Rightarrow \langle \mathbf{z}, \mathbf{v} - \mathbf{w} \rangle = D$.
- Let $\epsilon_1, \epsilon_2, \ldots, \epsilon_T$ be i.i.d. random variables such that $\Pr[\epsilon_t = 1] = \Pr[\epsilon_t = -1] = 1/2$ for each t.
- We choose the losses $\mathbf{g}_t = L\epsilon_t \mathbf{z}$.

Proof of Theorem B (1/2)

- The approach:

  For any random variable $\mathbf{z}$ with domain $\mathcal{V}$ and any function f,

$$\sup_{\mathbf{x} \in V} f(\mathbf{x}) \geq E[f(\mathbf{z})].$$

- $\text{regret}_T = \max_{\mathbf{u} \in \mathcal{K}} \text{regret}_T(\mathbf{u})$.
- Let $\mathbf{v}, \mathbf{w} \in \mathcal{K}$ such that $\|\mathbf{v} - \mathbf{w}\| = D$.
- Let $\mathbf{z} := \frac{\mathbf{v} - \mathbf{w}}{\|\mathbf{v} - \mathbf{w}\|} \Rightarrow \langle \mathbf{z}, \mathbf{v} - \mathbf{w} \rangle = D$.
- Let $\epsilon_1, \epsilon_2, \ldots, \epsilon_T$ be i.i.d. random variables such that $\Pr[\epsilon_t = 1] = \Pr[\epsilon_t = -1] = 1/2$ for each t.
- We choose the losses $\mathbf{g}_t = L\epsilon_t \mathbf{z}$.
  - The cost at t: $\langle L\epsilon_t \mathbf{z}, \mathbf{x}_t \rangle$.
  - $\|g_t\| = \sqrt{L^2 \epsilon_t^2} \cdot \|\mathbf{z}\| \leq L$.

Proof of Theorem B (2/2)

$$
\begin{aligned}
\sup_{\mathbf{g}_1,\ldots,\mathbf{g}_T} \text{regret}_T &\geq E\left[\sum_{t=1}^{T} L\epsilon_t\langle \mathbf{z}, \mathbf{x}_t\rangle - \min_{\mathbf{u}\in\mathcal{K}} \sum_{t=1}^{T} L\epsilon_t\langle \mathbf{z}, \mathbf{u}\rangle\right] \\
&= E\left[-\min_{\mathbf{u}\in\mathcal{K}} \sum_{t=1}^{T} L\epsilon_t\langle \mathbf{z}, \mathbf{u}\rangle\right] = E\left[\max_{\mathbf{u}\in\mathcal{K}} \sum_{t=1}^{T} L\epsilon_t\langle \mathbf{z}, \mathbf{u}\rangle\right] \\
&\geq E\left[\max_{\mathbf{u}\in\{\mathbf{v},\mathbf{w}\}} \sum_{t=1}^{T} L\epsilon_t\langle \mathbf{z}, \mathbf{u}\rangle\right] \\
&= E\left[\frac{1}{2}\sum_{t=1}^{T} L\epsilon_t\langle \mathbf{z}, \mathbf{v}+\mathbf{w}\rangle + \frac{1}{2}\left|\sum_{t=1}^{T} L\epsilon_t\langle \mathbf{z}, \mathbf{v}-\mathbf{w}\rangle\right|\right] \\
&\geq \frac{L}{2}E\left[\left|\sum_{t=1}^{T} \epsilon_t\langle \mathbf{z}, \mathbf{v}-\mathbf{w}\rangle\right|\right] = \frac{LD}{2}E\left[\left|\sum_{t=1}^{T} \epsilon_t\right|\right] \\
&\geq \frac{\sqrt{2}LD\sqrt{T}}{4}. \quad \text{(by Khintchine inequality)}
\end{aligned}
$$

Exercise

Prove the last inequality by Khintchine inequality.

Remark on the differentiability assumption

- The differentiability for loss function f_t is quite strong.

Remark on the differentiability assumption

- The differentiability for loss function f_t is quite strong.
- There are losses that are NOT differentiable.

Remark on the differentiability assumption

- The differentiability for loss function f_t is quite strong.
- There are losses that are NOT differentiable.
  - $f_t(x) = |x - 10|$, for $x \in \mathbb{R}$.

Remark on the differentiability assumption

- The differentiability for loss function f_t is quite strong.
- There are losses that are NOT differentiable.
  - $f_t(x) = |x - 10|$, for $x \in \mathbb{R}$.
  - $f_t(\mathbf{x}) = \max(1 - \langle \mathbf{z}, \mathbf{x} \rangle, 0)$ for $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$.

Remark on the differentiability assumption

- The differentiability for loss function f_t is quite strong.
- There are losses that are NOT differentiable.
  - $f_t(x) = |x - 10|$, for $x \in \mathbb{R}$.
  - $f_t(\mathbf{x}) = \max(1 - \langle \mathbf{z}, \mathbf{x} \rangle, 0)$ for $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$.
  - ReLU activation function: $f_t(x) = \max(x, 0)$.

Recall

Subgradient

For a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, $\mathbf{g} \in \mathbb{R}^d$ is a subgradient of f at $x \in \mathbb{R}^d$ if for all $\mathbf{y} \in \mathbb{R}^d$,

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

Recall

Subgradient

For a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, $\mathbf{g} \in \mathbb{R}^d$ is a subgradient of f at $x \in \mathbb{R}^d$ if for all $\mathbf{y} \in \mathbb{R}^d$,

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

- Note that the subgradient may NOT be unique!

Recall

Subgradient

For a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, $\mathbf{g} \in \mathbb{R}^d$ is a subgradient of f at $x \in \mathbb{R}^d$ if for all $\mathbf{y} \in \mathbb{R}^d$,

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

- Note that the subgradient may NOT be unique!
- We denote by $\partial f(\mathbf{x})$ the subdifferential at $\mathbf{x}$ which consists of all subgradients of f in $\mathbf{x}$.

Recall

Subgradient

For a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, $\mathbf{g} \in \mathbb{R}^d$ is a subgradient of f at $x \in \mathbb{R}^d$ if for all $\mathbf{y} \in \mathbb{R}^d$,

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

- Note that the subgradient may NOT be unique!
- We denote by $\partial f(\mathbf{x})$ the subdifferential at $\mathbf{x}$ which consists of all subgradients of f in $\mathbf{x}$.
- If f is convex, then $\partial f(\mathbf{x})$ turns out to be $\nabla f(\mathbf{x})$.

Examples and Useful Facts (1/5)

Theorem

Let $f_1, f_2, \ldots, f_m$ be convex functions on $\mathbb{R}^d$ and let $f = f_1 + f_2 + \ldots + f_m$. Then $\partial f(\mathbf{x}) \supseteq \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}) + \cdots + \partial f_m(\mathbf{x})$, for each $\mathbf{x} \in \mathbb{R}^d$.

- For any $\mathbf{z}$, define $\mathbf{g}_i \in \partial f_i(\mathbf{z})$ for $i = 1, 2, \ldots, m$.
- $f(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x}) \geq$

Examples and Useful Facts (1/5)

Theorem

Let $f_1, f_2, \ldots, f_m$ be convex functions on $\mathbb{R}^d$ and let $f = f_1 + f_2 + \ldots + f_m$. Then $\partial f(\mathbf{x}) \supseteq \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}) + \cdots + \partial f_m(\mathbf{x})$, for each $\mathbf{x} \in \mathbb{R}^d$.

- For any $\mathbf{z}$, define $\mathbf{g}_i \in \partial f_i(\mathbf{z})$ for $i = 1, 2, \ldots, m$.
- $f(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x}) \geq \sum_{i=1}^{m} (f_i(\mathbf{z}) + \langle \mathbf{g}_i, \mathbf{x} - \mathbf{z} \rangle)$

Examples and Useful Facts (1/5)

Theorem

Let $f_1, f_2, \ldots, f_m$ be convex functions on $\mathbb{R}^d$ and let $f = f_1 + f_2 + \ldots + f_m$. Then $\partial f(\mathbf{x}) \supseteq \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}) + \cdots + \partial f_m(\mathbf{x})$, for each $\mathbf{x} \in \mathbb{R}^d$.

- For any $\mathbf{z}$, define $\mathbf{g}_i \in \partial f_i(\mathbf{z})$ for $i = 1, 2, \ldots, m$.
- $f(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x}) \geq \sum_{i=1}^{m} (f_i(\mathbf{z}) + \langle \mathbf{g}_i, \mathbf{x} - \mathbf{z} \rangle) = f(\mathbf{z}) + \left\langle \sum_{i=1}^{m} \mathbf{g}_i, \mathbf{x} - \mathbf{z} \right\rangle.$

Examples and Useful Facts (1/5)

Theorem

Let $f_1, f_2, \ldots, f_m$ be convex functions on $\mathbb{R}^d$ and let $f = f_1 + f_2 + \ldots + f_m$. Then $\partial f(\mathbf{x}) \supseteq \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}) + \cdots + \partial f_m(\mathbf{x})$, for each $\mathbf{x} \in \mathbb{R}^d$.

- For any $\mathbf{z}$, define $\mathbf{g}_i \in \partial f_i(\mathbf{z})$ for $i = 1, 2, \ldots, m$.
- $f(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x}) \geq \sum_{i=1}^{m} (f_i(\mathbf{z}) + \langle \mathbf{g}_i, \mathbf{x} - \mathbf{z} \rangle) = f(\mathbf{z}) + \left\langle \sum_{i=1}^{m} \mathbf{g}_i, \mathbf{x} - \mathbf{z} \right\rangle.$

Examples and Useful Facts (2/5)

Subgradients of an Absolute Value Function

Let $f(x) = |x|$, then the subdifferential set $\partial f(x)$ is

$$\partial f(x) = \begin{cases} \{1\}, & x > 0, \\ \{-1\}, & x < 0, \\ [-1, 1], & x = 0. \end{cases}$$

Examples and Useful Facts (3/5)

Subgradients of the Hinge Loss

Consider $f : \mathbb{R}^d \mapsto \mathbb{R}$, such that $f(\mathbf{x}) = \max(1 - \langle \mathbf{z}, \mathbf{x} \rangle, 0)$ for $\mathbf{z} \in \mathbb{R}^d$. Then the subdifferential set $\partial f(\mathbf{x})$ is

$$\partial f(\mathbf{x}) = \begin{cases} \{\mathbf{0}\}, & \text{if } 1 - \langle \mathbf{z}, \mathbf{x} \rangle < 0 \\ \{-\alpha \mathbf{z} : \alpha \in [0, 1]\}, & \text{if } 1 - \langle \mathbf{z}, \mathbf{x} \rangle = 0 \\ \{-\mathbf{z}\}, & \text{otherwise} \end{cases}.$$

Examples and Useful Facts (4/5)

Subgradients of 2-Norm

Consider $f : \mathbb{R}^d \mapsto \mathbb{R}$, such that $f(\mathbf{x}) = \|\mathbf{x}\|_2$. Then the subdifferential set $\partial f(\mathbf{x})$ is

$$\partial f(\mathbf{x}) = \begin{cases} \mathbf{x}/\|\mathbf{x}\|_2, & \text{for } \mathbf{x} \neq \mathbf{0}, \\ \{\mathbf{z} : \|\mathbf{z}\|_2 \leq 1\}, & \text{for } \mathbf{x} = \mathbf{0}. \end{cases}$$

Examples and Useful Facts (5/5)

Lipschitz Continuity

A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is L-Lipschitz over a set V with respect to a norm $\|\cdot\|$ if $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in V$.

Theorem

Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a convex function. Then, f is L-Lipschitz in int dom(f) with respect to the L_2-norm if and only if

for all $\mathbf{x} \in$ int dom(f) and $\mathbf{g} \in \partial f(\mathbf{x})$, we have $\|\mathbf{g}\|_2 \leq L$.

- ⋆ The (sub-)gradient is also bounded by L!

A Recall for Conventional Continuity

Continuous Function

For a function $f : D \mapsto \mathbb{R}$, $D \subseteq \mathbb{R}$, f is continuous at $x_0 \in D$ if and only if

$$\forall \epsilon > 0, \exists \delta > 0, \text{ s.t. for all } x \in D, |x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \epsilon.$$

Proof

- Assume that f is L-Lipschitz, then

Proof

- Assume that f is L-Lipschitz, then $|f(\mathbf{x}) - f(\mathbf{y})| \leq$

Proof

- Assume that f is L-Lipschitz, then $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2$, for each $\mathbf{x}, \mathbf{y} \in \text{dom}(f)$.

Proof

- Assume that f is L-Lipschitz, then $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2$, for each $\mathbf{x}, \mathbf{y} \in \text{dom}(f)$.
- For small enough $\epsilon > 0$, let $\mathbf{y} = \mathbf{x} + \epsilon \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \text{int dom}(f)$.
  - $\|\mathbf{x} - \mathbf{y}\|_2 =$

Proof

- Assume that f is L-Lipschitz, then $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2$, for each $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$.
- For small enough $\epsilon > 0$, let $\mathbf{y} = \mathbf{x} + \epsilon \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathrm{int}\ \mathrm{dom}(f)$.
  - $\|\mathbf{x} - \mathbf{y}\|_2 = \epsilon$.

Proof

- Assume that f is L-Lipschitz, then $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2$, for each $\mathbf{x}, \mathbf{y} \in \text{dom}(f)$.
- For small enough $\epsilon > 0$, let $\mathbf{y} = \mathbf{x} + \epsilon \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \text{int dom}(f)$.
  - $\|\mathbf{x} - \mathbf{y}\|_2 = \epsilon$.
- Then,

$$L\epsilon = L\|\mathbf{x} - \mathbf{y}\|_2 \quad \geq \quad |f(\mathbf{x}) - f(\mathbf{y})|$$

Proof

- Assume that f is L-Lipschitz, then$|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2$, for each $\mathbf{x}, \mathbf{y} \in \text{dom}(f)$.
- For small enough $\epsilon > 0$, let $\mathbf{y} = \mathbf{x} + \epsilon\frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \text{int dom}(f)$.
  - $\|\mathbf{x} - \mathbf{y}\|_2 = \epsilon$.
- Then,

$$L\epsilon = L\|\mathbf{x} - \mathbf{y}\|_2 \quad \geq \quad |f(\mathbf{x}) - f(\mathbf{y})| \geq f(\mathbf{y}) - f(\mathbf{x})$$

Proof

- Assume that f is L-Lipschitz, then$|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2$, for each $\mathbf{x}, \mathbf{y} \in \text{dom}(f)$.
- For small enough $\epsilon > 0$, let $\mathbf{y} = \mathbf{x} + \epsilon \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \text{int dom}(f)$.
    - $\|\mathbf{x} - \mathbf{y}\|_2 = \epsilon$.
- Then,

$$\begin{aligned} L\epsilon = L\|\mathbf{x} - \mathbf{y}\|_2 &\geq |f(\mathbf{x}) - f(\mathbf{y})| \geq f(\mathbf{y}) - f(\mathbf{x}) \\ &\geq \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \end{aligned}$$

Proof

- Assume that f is L-Lipschitz, then$|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2$, for each $\mathbf{x}, \mathbf{y} \in \text{dom}(f)$.
- For small enough $\epsilon > 0$, let $\mathbf{y} = \mathbf{x} + \epsilon\frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \text{int dom}(f)$.
  - $\|\mathbf{x} - \mathbf{y}\|_2 = \epsilon$.
- Then,

$$
\begin{aligned}
L\epsilon = L\|\mathbf{x} - \mathbf{y}\|_2 &\geq |f(\mathbf{x}) - f(\mathbf{y})| \geq f(\mathbf{y}) - f(\mathbf{x}) \\
&\geq \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \\
&= \left\langle \mathbf{g}, \epsilon\frac{\mathbf{g}}{\|\mathbf{g}\|_2} \right\rangle
\end{aligned}
$$

Proof

- Assume that f is L-Lipschitz, then $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2$, for each $\mathbf{x}, \mathbf{y} \in \text{dom}(f)$.
- For small enough $\epsilon > 0$, let $\mathbf{y} = \mathbf{x} + \epsilon \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \text{int dom}(f)$.
    - $\|\mathbf{x} - \mathbf{y}\|_2 = \epsilon$.
- Then,

$$\begin{aligned} L\epsilon = L\|\mathbf{x} - \mathbf{y}\|_2 &\geq |f(\mathbf{x}) - f(\mathbf{y})| \geq f(\mathbf{y}) - f(\mathbf{x}) \\ &\geq \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \\ &= \left\langle \mathbf{g}, \epsilon \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \right\rangle = \epsilon \|\mathbf{g}\|_2. \end{aligned}$$

Projected Online Subgradient Descent

1. **Input:** convex set $\mathcal{K}$, T, $\mathbf{x}_1 \in \mathcal{K}$, 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. Set $\mathbf{g}_t \in \partial f_t(\mathbf{x}_t)$.
    3. Update and Project:

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

3. **end for**

Discussions