Online Learning

— The Multiplicative-Weight Update Algorithm

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

Listen to the experts?

- Let's say we have n experts.
- We want to make best use of the advices coming from the experts.

Listen to the experts?

- Let's say we have n experts.
- We want to make best use of the advices coming from the experts.
- The idea: at each time step, decide the probability distribution (i.e., weights) of the experts to follow their advice.
  - $\mathbf{x}_t = (\mathbf{x}_t(1), \mathbf{x}_t(2), \ldots, \mathbf{x}_t(n))$, where $\mathbf{x}_t(i) \in [0, 1]$ and $\sum_i \mathbf{x}_t(i) = 1$.

Listen to the experts?

- Let's say we have n experts.
- We want to make best use of the advices coming from the experts.
- The idea: at each time step, decide the probability distribution (i.e., weights) of the experts to follow their advice.
  - $\mathbf{x}_t = (\mathbf{x}_t(1), \mathbf{x}_t(2), \ldots, \mathbf{x}_t(n))$, where $\mathbf{x}_t(i) \in [0, 1]$ and $\sum_i \mathbf{x}_t(i) = 1$.
- The loss of following expert i at time t: $\ell_t(i)$.
- The expected loss of the algorithm at time t:

$$\langle \mathbf{x}_t, \ell_t \rangle = \sum_{i=1}^{n} \mathbf{x}_t(i) \ell_t(i).$$

The regret of listening to the experts...

$$\text{regret}_T = \sum_{t=1}^{T} \langle \mathbf{x}_t, \ell_t \rangle - \min_i \sum_{t=1}^{T} \ell_t(i).$$

- The set of feasible solutions $K = \Delta \subseteq \mathbb{R}^n$, probability distributions over $\{1, \ldots, n\}$.
- $f_t(\mathbf{x}) = \sum_i \mathbf{x}(i)\ell_t(i)$: linear function.

⋆ Assume that $|\ell_t(i)| \leq 1$ for all t and i.
  - Normalized.

The regret of listening to the experts...

$$\text{regret}_T = \sum_{t=1}^{T} \langle \mathbf{x}_t, \ell_t \rangle - \min_i \sum_{t=1}^{T} \ell_t(i).$$

- The set of feasible solutions $K = \Delta \subseteq \mathbb{R}^n$, probability distributions over $\{1, \ldots, n\}$.
- $f_t(\mathbf{x}) = \sum_i \mathbf{x}(i)\ell_t(i)$: linear function.

⋆ Assume that $|\ell_t(i)| \leq 1$ for all t and i.
  - Normalized.

- In fact, we claim that (exercise!)

$$\min_i \sum_{t=1}^{T} \ell_t(i) = \min_{\mathbf{x}} \sum_{t=1}^{T} \langle \mathbf{x}, \ell_t \rangle.$$

The MWU Algorithm

- The spirit: "Hedge".
- Well-known and frequently rediscovered.

The MWU Algorithm

- The spirit: "Hedge".
- Well-known and frequently rediscovered.

Multiplicative Weight Update (MWU)

- Maintain a vector of weights $\mathbf{w}_t = (\mathbf{w}_t(1), \ldots, \mathbf{w}_t(n))$ where $\mathbf{w}_1 := (1, 1, \ldots, 1)$.
- Update the weights at time t by
  - $\mathbf{w}_t(i) := \mathbf{w}_{t-1}(i) \cdot e^{-\beta \ell_{t-1}(i)}$.
  - $\mathbf{x}_t := \frac{\mathbf{w}_t(i)}{\sum_{j=1}^{n} \mathbf{w}_t(j)}$.

β: a parameter which will be optimized later.

The MWU Algorithm

- The spirit: "Hedge".
- Well-known and frequently rediscovered.

Multiplicative Weight Update (MWU)

- Maintain a vector of weights $\mathbf{w}_t = (\mathbf{w}_t(1), \ldots, \mathbf{w}_t(n))$ where $\mathbf{w}_1 := (1, 1, \ldots, 1)$.
- Update the weights at time t by
  - $\mathbf{w}_t(i) := \mathbf{w}_{t-1}(i) \cdot e^{-\beta \ell_{t-1}(i)}$.
  - $\mathbf{x}_t := \frac{\mathbf{w}_t(i)}{\sum_{j=1}^{n} \mathbf{w}_t(j)}$.

β: a parameter which will be optimized later.

The weight of expert i at time t: $e^{-\beta \sum_{k=1}^{t-1} \ell_k(i)}$.

MWU is of no-regret

Theorem 1 (MWU is of no-regret)

Assume that $|\ell_t(i)| \leq 1$ for all t and i. For $\beta \in (0, 1/2)$, the regret of MWU after T steps is bounded as

$$\text{regret}_T \leq \beta \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{x}_t(i)\ell_t^2(i) + \frac{\ln n}{\beta} \leq \beta T + \frac{\ln n}{\beta}.$$

In particular, if $T > 4 \ln n$, then

$$\text{regret}_T \leq 2\sqrt{T \ln n}$$

by setting $\beta = \sqrt{\frac{\ln n}{T}}$.

Proof of Theorem 1

Let $W_t := \sum_{i=1}^{n} \mathbf{w}_t(i)$.

- The total weight at time t.

The idea:

- If the algorithm incurs a large loss after T steps, then W_{T+1} is small.
- And, if W_{T+1} is small, then even the best expert performs quite badly.

Proof of Theorem 1

Let $W_t := \sum_{i=1}^{n} \mathbf{w}_t(i)$.

- The total weight at time t.

The idea:

- If the algorithm incurs a large loss after T steps, then W_{T+1} is small.
- And, if W_{T+1} is small, then even the best expert performs quite badly.

Let $L^* := \min_i \sum_{t=1}^{T} \ell_t(i)$.

- The cumulative loss of the "best" expert.

The proof (contd.)

Lemma 1 (W_{T+1} is SMALL $\Rightarrow$ L^* is LARGE)

$W_{T+1} \geq e^{-\beta L^*}$.

Proof.

Let $j = \arg\min L^* = \arg\min_i \sum_{t=1}^{T} \ell_t(i)$.

$$W_{T+1} = \sum_{i=1}^{n} e^{-\beta \sum_{t=1}^{T} \ell_t(i)} \geq e^{-\beta \sum_{t=1}^{T} \ell_t(j)} = e^{-\beta L^*}.$$

□

The proof (contd.)

Lemma 2 (MWU brings large loss ⇒ W_{T+1} is SMALL)

$$W_{T+1} \le n \prod_{t=1}^{n} (1 - \beta\langle \mathbf{x}_t, \ell_t \rangle + \beta^2 \langle \mathbf{x}_t, \ell_t^2 \rangle),$$

Proof.

Note: $W_1 = n$.

$$\frac{W_{t+1}}{W_t} = \sum_{i=1}^{n} \frac{\mathbf{w}_{t+1}(i)}{W_t} = \sum_{i=1}^{n} \frac{\mathbf{w}_t(i) \cdot e^{-\beta \ell_t(i)}}{W_t}$$

The proof (contd.)

Lemma 2 (MWU brings large loss $\Rightarrow W_{T+1}$ is SMALL)

$$W_{T+1} \leq n \prod_{t=1}^{n} (1 - \beta \langle \mathbf{x}_t, \ell_t \rangle + \beta^2 \langle \mathbf{x}_t, \ell_t^2 \rangle),$$

Proof.

Note: $W_1 = n$.

$$\begin{aligned} \frac{W_{t+1}}{W_t} &= \sum_{i=1}^{n} \frac{\mathbf{w}_{t+1}(i)}{W_t} = \sum_{i=1}^{n} \frac{\mathbf{w}_t(i) \cdot e^{-\beta \ell_t(i)}}{W_t} = \sum_{i=1}^{n} \mathbf{x}_t(i) \cdot e^{-\beta \ell_t(i)} \\ &\leq \sum_{i=1}^{n} \mathbf{x}_t(i) \cdot (1 - \beta \ell_t(i) + \beta^2 \ell_t^2(i)) \end{aligned}$$

The proof (contd.)

Lemma 2 (MWU brings large loss $\Rightarrow W_{T+1}$ is SMALL)

$$W_{T+1} \le n \prod_{t=1}^{n} (1 - \beta\langle \mathbf{x}_t, \ell_t \rangle + \beta^2 \langle \mathbf{x}_t, \ell_t^2 \rangle),$$

Proof.

Note: $W_1 = n$.

$$\begin{aligned} \frac{W_{t+1}}{W_t} &= \sum_{i=1}^{n} \frac{\mathbf{w}_{t+1}(i)}{W_t} = \sum_{i=1}^{n} \frac{\mathbf{w}_t(i) \cdot e^{-\beta \ell_t(i)}}{W_t} = \sum_{i=1}^{n} \mathbf{x}_t(i) \cdot e^{-\beta \ell_t(i)} \\ &\le \sum_{i=1}^{n} \mathbf{x}_t(i) \cdot (1 - \beta \ell_t(i) + \beta^2 \ell_t^2(i)) \\ &= 1 - \beta \langle \mathbf{x}_t, \ell_t \rangle + \beta^2 \langle \mathbf{x}_t, \ell_t^2 \rangle \end{aligned}$$

The proof (contd.)

Lemma 2 (MWU brings large loss $\Rightarrow W_{T+1}$ is SMALL)

$$W_{T+1} \le n \prod_{t=1}^{n} \left(1 - \beta\langle \mathbf{x}_t, \ell_t \rangle + \beta^2 \langle \mathbf{x}_t, \ell_t^2 \rangle\right),$$

Proof.

Note: $W_1 = n$.

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{i=1}^{n} \frac{\mathbf{w}_{t+1}(i)}{W_t} = \sum_{i=1}^{n} \frac{\mathbf{w}_t(i) \cdot e^{-\beta \ell_t(i)}}{W_t} = \sum_{i=1}^{n} \mathbf{x}_t(i) \cdot e^{-\beta \ell_t(i)} \\
&\le \sum_{i=1}^{n} \mathbf{x}_t(i) \cdot (1 - \beta \ell_t(i) + \beta^2 \ell_t^2(i)) \\
&= 1 - \beta\langle \mathbf{x}_t, \ell_t \rangle + \beta^2 \langle \mathbf{x}_t, \ell_t^2 \rangle \le e^{-\beta\langle \mathbf{x}_t, \ell_t \rangle + \beta^2 \langle \mathbf{x}_t, \ell_t^2 \rangle}.
\end{aligned}$$

□

The proof (contd.)

Lemma 2 (MWU brings large loss $\Rightarrow W_{T+1}$ is SMALL)

$$W_{T+1} \le n \prod_{t=1}^{n} e^{-\beta\langle x_t, \ell_t\rangle + \beta^2 \langle x_t, \ell_t^2\rangle}.$$

Proof.

Note: $W_1 = n$.

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{i=1}^{n} \frac{\mathbf{w}_{t+1}(i)}{W_t} = \sum_{i=1}^{n} \frac{\mathbf{w}_t(i) \cdot e^{-\beta \ell_t(i)}}{W_t} = \sum_{i=1}^{n} \mathbf{x}_t(i) \cdot e^{-\beta \ell_t(i)} \\
&\le \sum_{i=1}^{n} \mathbf{x}_t(i) \cdot (1 - \beta \ell_t(i) + \beta^2 \ell_t^2(i)) \\
&= 1 - \beta\langle \mathbf{x}_t, \ell_t\rangle + \beta^2 \langle \mathbf{x}_t, \ell_t^2\rangle \le e^{-\beta\langle \mathbf{x}_t, \ell_t\rangle + \beta^2 \langle \mathbf{x}_t, \ell_t^2\rangle}.
\end{aligned}$$

□

The proof (contd.)

Hence

$$\ln W_{T+1} \le \ln n - \left(\sum_{i=1}^{T} \beta \langle \ell_t, \mathbf{x}_t \rangle \right) + \left(\sum_{i=1}^{T} \beta^2 \langle \ell_t^2, \mathbf{x}_t \rangle \right)$$

and $\ln W_{T+1} \ge -\beta L^*$ (by Lemma 1).

The proof (contd.)

Hence

$$\ln W_{T+1} \le \ln n - \left(\sum_{i=1}^{T} \beta \langle \ell_t, \mathbf{x}_t \rangle \right) + \left(\sum_{i=1}^{T} \beta^2 \langle \ell_t^2, \mathbf{x}_t \rangle \right)$$

and $\ln W_{T+1} \ge -\beta L^*$ (by Lemma 1).

Thus,

$$\left(\sum_{t=1}^{T} \langle \ell_t, \mathbf{x}_t \rangle \right) - L^* \le \frac{\ln n}{\beta} + \beta \sum_{t=1}^{T} \langle \ell_t^2, \mathbf{x}_t \rangle .$$

The proof (contd.)

Hence

$$\ln W_{T+1} \leq \ln n - \left(\sum_{i=1}^{T} \beta\langle \ell_t, \mathbf{x}_t\rangle\right) + \left(\sum_{i=1}^{T} \beta^2\langle \ell_t^2, \mathbf{x}_t\rangle\right)$$

and $\ln W_{T+1} \geq -\beta L^*$ (by Lemma 1).

Thus,

$$\left(\sum_{t=1}^{T} \langle \ell_t, \mathbf{x}_t\rangle\right) - L^* \leq \frac{\ln n}{\beta} + \beta \sum_{t=1}^{T} \langle \ell_t^2, \mathbf{x}_t\rangle.$$

Take $\beta = \sqrt{\frac{\ln n}{T}}$, we have $\text{regret}_T \leq 2\sqrt{T \ln n}$.

The proof (contd.)

Hence

$$\ln W_{T+1} \leq \ln n - \left(\sum_{i=1}^{T} \beta \langle \ell_t, \mathbf{x}_t \rangle\right) + \left(\sum_{i=1}^{T} \beta^2 \langle \ell_t^2, \mathbf{x}_t \rangle\right)$$

and $\ln W_{T+1} \geq -\beta L^*$ (by Lemma 1).

Thus,

$$\left(\sum_{t=1}^{T} \langle \ell_t, \mathbf{x}_t \rangle\right) - L^* \leq \frac{\ln n}{\beta} + \beta \sum_{t=1}^{T} \langle \ell_t^2, \mathbf{x}_t \rangle.$$

Take $\beta = \sqrt{\frac{\ln n}{T}}$, we have $\text{regret}_T \leq 2\sqrt{T \ln n}$.

Note: $\sum_{i=1}^{n} \mathbf{x}_t(i) = 1$ and $0 \leq \ell_t^2(i) \leq 1$.

Discussions