Online Optimization, Learning, and Games (O2LG)
Lesson 6: Online Learning in Discrete Time

Vinh Thanh Ho*, Panayotis Mertikopoulos

*Faculté des Sciences et Techniques
Université de Limoges
vinh-thanh.ho@unilim.fr

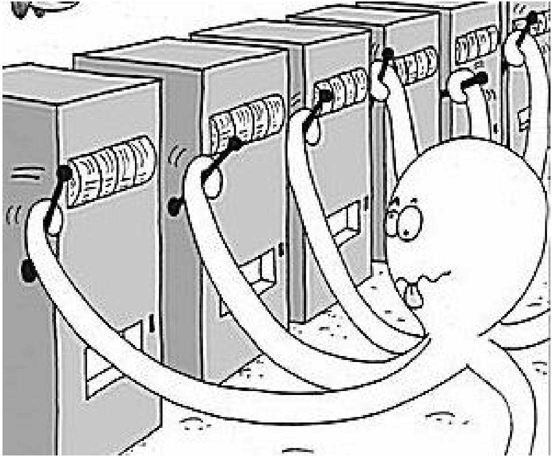
ACSYON
Master of Applied Mathematics

Table of Contents

1. Multi-armed bandits
2. Regret in discrete time
3. Exponential weights in discrete time

Multi-armed bandits

Robbins' multi-armed bandit problem (Robbins 1952): how to play in a (rigged) casino?



Online learning in discrete time

Sequence of events — discrete time

Require: set of actions $\mathcal{A} = \{1, \dots, A\}$, sequence of payoff vectors v_n, $n = 1, 2, \dots$
for all $n = 1, 2, \dots$ **do**
Choose mixed strategy $x_n \in \mathcal{X} := \Delta(\mathcal{A})$.
Play action $a_n \sim x_n$.
Encounter payoff vector v_n and receive payoff $u_n(a_n) = v_{a_n,n}$.
end for

Features: *discrete* time, *single* player, and *exogenous* payoffs.

Three types of **feedback** (from best to worst):

- Full, exact information: observe entire payoff vector v_n.
- Full, inexact information: observe noisy estimate of v_n.
- Partial information / Bandit: only chosen component $u_n(a_n) = v_{a_n,n}$.

Feedback

The oracle model

A stochastic first-order oracle (SFO) model of v_n is a random vector of the form

$$\hat{v}_n = v_n + U_n + b_n, \tag{SFO}$$

where U_n is zero-mean and $b_n = \mathbb{E}[\hat{v}_n \mid \mathcal{F}_n] - v(x_n)$ is the bias of $\hat{v}_n$.

Assumptions 1

- Bias: $\|b_n\| \leq B_n$.
- Variance: $\mathbb{E}[\|U_n\|^2 \mid \mathcal{F}_n] \leq \sigma_n^2$.
- Second moment: $\mathbb{E}[\|\hat{v}_n\|^2 \mid \mathcal{F}_n] \leq M_n^2$.

Recall: Reconstructing payoff vectors

Definition 1 (Importance-weighted estimator)

Fix a payoff vector $v \in \mathbb{R}^{\mathcal{A}}$ and a probability distribution P on $\mathcal{A}$. Then, for a given $a \in \mathcal{A}$, the importance-weighted estimator of v_a relative to P is the **random variable**

$$\hat{v}_a = \frac{v_a}{P_a} \mathbb{1}_a = \begin{cases} \dfrac{v_a}{P_a} & \text{if } a \text{ is drawn,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Statistical properties of (1) in IWE

- Unbiasedness: $\mathbb{E}[\hat{v}_a] = v_a$.
- Second moment: $\mathbb{E}[\hat{v}_a^2] = \dfrac{v_a^2}{P_a}$.

Table of Contents

1. Multi-armed bandits
2. Regret in discrete time
3. Exponential weights in discrete time

Regret

The agent's regret in discrete time:

Realized regret: $$\overline{\mathrm{Reg}}(T) = \max_{a\in\mathcal{A}} \sum_{n=1}^{T} [u_n(a) - u_n(a_n)].$$

Mean regret: $$\mathrm{Reg}(T) = \max_{p\in\mathcal{X}} \sum_{n=1}^{T} [u_n(p) - u_n(x_n)] = \max_{p\in\mathcal{X}} \underbrace{\sum_{n=1}^{T} \langle v_n, p - x_n \rangle}_{\mathrm{Reg}_p(T)} .$$

Table of Contents

1. Multi-armed bandits
2. Regret in discrete time
3. Exponential weights in discrete time

The exponential weights algorithm

Basic idea:

- Score actions by aggregating oracle feedback signals.
- Choose an action with probability exponentially proportional to its score.
- Rinse / repeat.

Algorithm 1 Exponential Weights with oracle feedback (ExpWeight)

Require: set of actions $\mathcal{A}$, sequence of payoff vectors v_n, and SFO feedback $\hat{v}_n$, $n = 1, 2, \dots$

  Initialize: $y_1 \in \mathbb{R}^{\mathcal{A}}$.

  for all $n = 1, 2, \dots$ **do**

    set $x_n \leftarrow \Lambda(y_n)$. ▷ mixed strategy

    play $a_n \sim x_n$ and receive $v_{a_n,n}$. ▷ choose action / get payoff

    observe $\hat{v}_n \in \mathbb{R}^{\mathcal{A}}$. ▷ receive feedback

    set $y_{n+1} \leftarrow y_n + \gamma_n \hat{v}_n$. ▷ update scores

  end for

Regret analysis

Theorem 1 (Auer et al. 1995)

- *Assume:*
  - *SFO sequence $(\hat{v}_n)_n$ unbiased and bounded in mean square* ($B_n = 0$, $\sup_n M_n < M$).
  - $\gamma = M^{-1}\sqrt{(2\log A)/T}$.
- *Then, for all $p \in \mathcal{X}$,* ExpWeight *enjoys the bound*

$$\mathrm{Reg}(T) \le M\sqrt{2\log A \cdot T} = \mathcal{O}(\sqrt{T}).$$

Mimic the continuous-time case:

- Use a constant $\gamma_n \equiv \gamma$.
- Fix the comparator $p \in \mathcal{X}$ and consider the Fenchel coupling:

$$F_n := F(p, y_n) = \sum_{a\in\mathcal{A}} p_a \log p_a + \log \sum_{a\in\mathcal{A}} \exp(y_{a,n}) - \langle y_n, p \rangle.$$

Energy inequality

$$F_n := F(p, y_n) = \sum_{a \in \mathcal{A}} p_a \log p_a + \log \sum_{a \in \mathcal{A}} \exp(y_{a,n}) - \langle y_n, p \rangle.$$

Energy inequality

$$F_{n+1} \leq F_n + \gamma \langle \hat{v}_n, x_n - p \rangle + \mathcal{O}(\gamma^2). \tag{2}$$

Lemma 1

For all $y, w \in \mathbb{R}^{\mathcal{A}}$, we have: $\log \sum_{a \in \mathcal{A}} \exp(y_a + w_a) \leq \log \sum_{a \in \mathcal{A}} \exp(y_a) + \langle \Lambda(y), w \rangle + \frac{1}{2} \|w\|_\infty^2$.

Task 1

Prove and use Lemma 1 to establish the energy inequality (2).

Regret analysis

- By simplifying and taking expectations, we get

$$\mathrm{Reg}_p(T) \leq \frac{F_1}{\gamma} + \mathcal{O}(\gamma).$$

- Balancing right-hand-side with the step-size γ (defined in the assumption) implies that

$$\mathrm{Reg}(T) \leq M\sqrt{2\log A \cdot T} = \mathcal{O}(\sqrt{T}). \tag{3}$$

Task 2

Prove the regret's bound (3).

Regret of ExpWeight

$$\mathrm{Reg}(T) = \mathcal{O}(\sqrt{T}).$$

Remarks:

- This bound is tight in T. # Abernethy et al. 2008
- Logarithmic dependence on A. # Can deal with exponentially many arms!!
- Cannot achieve $\mathcal{O}(1)$ regret as in continuous time.

Summary

This lesson

- Online learning in discrete time
- $\mathcal{O}(\sqrt{T})$ regret in discrete time

Next lesson

- Online convex optimization

References

[1] Jacob Abernethy et al. Optimal Stragies and Minimax Lower Bounds for Online Convex Games. In: Jan. 2008, pp. 415–424 (cited at slide -1).

[2] P. Auer et al. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In: *Proceedings of IEEE 36th Annual Foundations of Computer Science*. 1995, pp. 322–331 (cited at slide -4).

[3] Herbert Robbins. Some aspects of the sequential design of experiments. In: *Bulletin of the American Mathematical Society* 58.5 (1952), pp. 527–535 (cited at slide -12).