Online Learning in Games

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 4: Online learning with bandit feedback

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-735 (Spring 2024)

License Information for Online Learning in Games Slides

- This work is released under a Creative Commons License with the following terms:
- **Attribution**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.
- **Non-Commercial**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor's permission.
- **Share Alike**
  - The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor's work.
- Full Text of the License

Outline of this lecture

Online decision making: Feedback types

The sequence of events in an online decision problem

Let us define the set of actions as $\mathcal{A} = \{1, 2, \ldots, K\}$, and the sequence of cost vectors c_t (typically bounded). At each round $t = 1, \ldots, T$, where T is the time horizon,

- A learner selects an $x^t \in \Delta(\mathcal{A})$ (i.e., a distribution over the actions).
- The learner plays an action $a^t \in \mathcal{A} \sim x^t$.
- An adversary selects a cost vector $c_t = (c_{t,1}, \ldots, c_{t,K})$.
- The learner suffers the cost $f_t(a^t) = c_{t,a^t}$.

Types of feedback (from best to worst)

- *Full information*: The learner observes the (whole) cost vector c_t.
- *Exact and/or inexact cost information*: The learner observes a noisy version of the (whole) cost vector c_t.
- *Bandit information*: The learner observes only the cost at time t: $f_t(a^t) = c_{t,a^t}$.

This lecture: Online decision processes with bandit feedback

Objectives:

- Establish appropriate algorithmic solution methods to online decision problems.
- Examine how **regret minimization** is influenced by the learner's **limited** feedback.
- Establish lower-bounds and the optimality of the respective methods.

The multi-armed bandit problem

The (MAB) formulation

Let us define the set of actions as $\mathcal{A} = \{1, 2, \ldots, K\}$, and the sequence of cost vectors c_t (typically bounded).

At each round $t = 1, \ldots, T$, where T is the time horizon,

- A learner selects an $x^t \in \Delta(\mathcal{A})$ (i.e., a distribution over the actions).
- The learner plays an action $a^t \in \mathcal{A} \sim x^t$.
- An adversary selects a cost vector $c_t = (c_{t,1}, \ldots, c_{t,K})$.
- The learner suffers a cost $f_t(a^t) = c_{t,a^t}$.

Remark:

- Note that $f_t(a^t) = c_{t,a^t}$ is the learner's only feedback.

Expert advice vs Multi-armed bandit (MAB)

- In the MAB setting, the learner does not have access to the cost vector c_t at time t!

Expert advice problem

Set of actions $\mathcal{A} = \{1, 2, \ldots, K\}$, sequence of cost vectors c_t. At each round $t = 1, \ldots, T$, where T is the time horizon,

- A learner selects an $x^t \in \Delta(\mathcal{A})$.
- The learner plays an action $a^t \in \mathcal{A} \sim x^t$.
- An adversary selects a $c_t = (c_{t,1}, \ldots, c_{t,K})$.
- The learner suffers a cost $\langle c_t, x^t \rangle$ and receives the whole cost vector c_t.

(MAB)

Set of actions $\mathcal{A} = \{1, 2, \ldots, K\}$, sequence of cost vectors c_t. At each round $t = 1, \ldots, T$, where T is the time horizon,

- A learner selects an $x^t \in \Delta(\mathcal{A})$.
- The learner plays an action $a^t \in \mathcal{A} \sim x^t$.
- An adversary selects a $c_t = (c_{t,1}, \ldots, c_{t,K})$.
- The learner suffers a cost $f_t(a^t) = c_{t,a^t}$ and this is their *only* feedback.

Remarks:

- Expert advice problem <> first-order optimization
- Multi-armed bandit problem <> (stochastic) zero-th order optimization

Towards building a solution method for MAB

Step I: Construct an appropriate stochastic oracle

We define an appropriate unbiased estimator for the cost vector c_t for every round $t = 1, \dots, T$.

Step II: Construct an appropriate algorithmic template

Given the specific stochastic oracle, we define an appropriate recursive update, which will guarantee low regret.

Constructing an oracle

- Intuition: Mimicking SGD, we ideally need an estimator with the following statistical assumptions:

Assumption (Unbiasedness)

For all $a \in \mathcal{A}$ and $t = 1, \dots T$, we would need the following

$$\mathbb{E}[\hat{c}_t] = c_t.$$

Assumption (Bounded second moment)

For all $t = 1, \dots T$, we would need the following

$$\mathbb{E}[\|\hat{c}_t\|_\infty^2] \leq \sigma^2$$

The estimator

- **Main goal**: We want from a "single-action" feedback to estimate the whole cost vector!

Importance weighted estimators (IWE)

Given a cost vector c_t and a probability distribution $x^t \in \Delta(\mathcal{A})$, we define the importance weighted distribution of $\hat{c}_t = (\hat{c}_{t,a})_{a \in \mathcal{A}}$ as follows:

$$\hat{c}_{t,a} = \frac{\mathbf{1}_a}{x_a^t} c_{t,a} := \begin{cases} \frac{c_{t,a}}{x_a^t} & \text{if } a \text{ is drawn } (a = a'); \\ 0 & \text{otherwise } (a \neq a'); \end{cases} \tag{IWE}$$

where $\mathbf{1}_a$ is the indicator function for the action a as defined above.

Statistical properties of IWE: Part I

- A Natural question: Are unbiasedness / bounded second moment satisfied by IWE?

Unbiasedness

The IWE is an unbiased estimator of the cost vector c_t. In particular, for all $a \in \mathcal{A}$, it holds that

$$\mathbb{E}[\hat{c}_{t,a}] = c_{t,a}.$$

Proof

For all $a \in \mathcal{A}$, it holds that

$$\mathbb{E}[\hat{c}_{t,a}] = \sum_{a' \in \mathcal{A}} x^t_{a'} \frac{\mathbf{1}_a}{x^t_a} c_{t,a} = x^t_a \frac{1}{x^t_a} c_{t,a} = c_{t,a},$$

and hence the result follows.

Remark: - The analogy with the zero-th order optimization is clear.

Statistical properties of IWE: Part II

- After establishing the unbiasedness, we now seek to upper-bound the second moment.

Lemma (Second moment bound)

The second moment of IWE is of order $\mathcal{O}(1/x_a^t)$. In particular, for all $a \in \mathcal{A}$, it holds that

$$\mathbb{E}[\|\hat{c}_{t,a}\|_\infty^2] = \frac{c_{t,a}^2}{x_a^t}.$$

Remark:
- IWE does not have bounded second-order moment since it "explodes" when $x_a \to 0$.

Exercise:
- Show that the above holds.

The mirror descent (MD) template

- We present the key-elements of the MD template here.

Building blocks of MD

- Set a regularization function h, typically assumed to be K- strongly convex. That is,

$$h(x) \geq h(y) + \langle \nabla h(y), y - x \rangle + \frac{\mu}{2}\|x - y\|^2, \forall x, y \in \mathcal{X}.$$

- Define the so-called mirror map as follows:

$$\mathcal{Q}(v) = \arg\min_{x \in \mathcal{X}} \{\langle v, x \rangle + h(x)\},$$

where $\mathcal{X}$ denotes the feasible domain.

Projected vs mirrored updates

Projected Gradient Descent (PGD)

Set step-size policy γ_t, sequence of vectors v_t generated by the respective oracle and $\mathcal{X}$ feasible domain.

- *Aggregate* oracle's feedback $Y_{t+1} = Y_t - \gamma_t v_t$.
- *Update*:

$$x^{t+1} = \arg\min_{x \in \mathcal{X}} \{\langle Y_{t+1}, x \rangle + 1/2 \|x\|^2\}$$

Mirror Descent (MD) Shalev-Shwartz [7]

Set step-size policy γ_t, sequence of vectors v_t generated by the respective oracle and $\mathcal{X}$ feasible domain.

- *Aggregate* oracle's feedback $Y_{t+1} = Y_t - \gamma_t v_t$.
- *Update*:

$$x^{t+1} = \arg\min_{x \in \mathcal{X}} \{\langle Y_{t+1}, x \rangle + h(x)\} = \mathcal{Q}(Y_{t+1})$$

Remarks:

- Crucial difference: The choice of a generic regularizer h.
- Note that for different choices we obtain different algorithm.
- Important example, for $h(x) = 1/2\|x\|^2$, then (PGD) and (MD) coincide.

Overview of (MD) variants

Dual Averaging Nesterov [6]

Set step-size policy γ_t, sequence of vectors v_t generated by the respective oracle and $\mathcal{X}$ feasible domain.

- *Aggregate* oracle's feedback $Y_{t+1} = Y_t - v_t$.
- *Update*:

$$x^{t+1} = \mathcal{Q}(\gamma_{t+1} Y_{t+1})$$

Mirror Descent (MD) Shalev-Shwartz [7]

Set step-size policy γ_t, sequence of vectors v_t generated by the respective oracle and $\mathcal{X}$ feasible domain.

- *Aggregate* oracle's feedback $Y_{t+1} = Y_t - \gamma_t v_t$.
- *Update*:

$$x^{t+1} = \mathcal{Q}(Y_{t+1})$$

Remark:

- The crucial difference between (DA) and (MD) is the post vs pre- multiplication of the dual sequence.
- In (DA) the dual sequence enters always with *no* weights and the learning rate applies after the aggregation.
- In (MD) the dual sequence is weighted and then we take the aggregation.

Energy inequality of (MD)

Fenchel coupling, Mertikopoulos, Zhou [5]

We define the so-called Fenchel coupling which serves in the sequel as the appropriate Lyapunov function:

$$F(x,y) = h(x) + h^*(y) - \langle y, x\rangle$$

with $h^*(v) = \sup_{x\in\mathcal{X}}\{\langle v, x\rangle - h(x)\}$ (dubbed as Fenchel conjugate).

Energy inequality

Assume that x^t are the iterates generated by the (MD) algorithm. Then, the following inequality holds:

$$F_{t+1} \leq F_t - \gamma\langle v_t, x^t - p\rangle + \frac{1}{2}\gamma^2\|v_t\|_\infty^2$$

or equivalently after rearranging:

$$\langle v_t, x^t - p\rangle \leq \frac{F_t - F_{t+1}}{\gamma} + \frac{\gamma}{2}\|v_t\|_\infty^2$$

Proof of energy inequality: Part I

In order to prove the energy inequality, we need the following technical lemmas:

Three-point identity, Antonakopoulos et al [2]

Let h be a regularization function and $x = \mathcal{Q}(y)$ Then, by fixing $x^* \in \mathcal{X}$ and $y, y^+ \in \mathcal{X}^*$, it holds that

$$F(x^*, y^+) = F(x^*, y) + F(x, y^+) + \langle y^+ - y, x - x^* \rangle$$

Norm compatibility, Mertikopoulos, Zhou [5]

Assume that h is μ- strongly convex, i.e.,

$$h(x) \geq h(y) + \langle \nabla h(y), y - x \rangle + \frac{\mu}{2}\|x - y\|^2,$$

then, the following holds:

$$F(x, y) \geq \frac{\mu}{2}\|\mathcal{Q}(y) - x\|^2.$$

Proof of energy inequality: Part II

- We now turn our attention to the actual proof of the generic (MD) energy inequality:

Proof

We have:

$$\begin{aligned}
\langle v_t, x^t - p\rangle &= \langle v_t, x^{t+1} - p\rangle + \langle v_t, x^t - x^{t+1}\rangle \\
&= \frac{1}{\gamma_t}\langle Y_t - Y_{t+1}, x^{t+1} - p\rangle + \langle v_t, x^t - x^{t+1}\rangle \\
&= \frac{1}{\gamma_t}\left[F_t - F_{t+1} - F(x^{t+1}, Y_t)\right] + \langle v_t, x^t - x^{t+1}\rangle \\
&\leq \frac{F_t - F_{t+1}}{\gamma_t} + \left[F(x^{t+1}, Y_t) - \frac{K}{2}\|x^t - x^{t+1}\|^2\right] + \frac{\gamma_t}{2K}\|v_t\|^2 \quad \leq \frac{F_t - F_{t+1}}{\gamma_t} + \frac{\gamma_t}{2K}\|v_t\|^2
\end{aligned}$$

and the result follows.

The (EXP3) algorithm

Main ingredients

(EXP3) relies on the (MD) template with

- the *entropic regularizer* $h(p) = \sum_{a \in \mathcal{A}} p_a \log p_a$
- the (IWE) as an feedback vector, ie $v_t = \hat{c}_t$.

The EXP3 algorithm

- Require set of actions $\mathcal{A}$, sequence of cost vectors $c_t \in [0,1]^{\mathcal{A}}$.
- Initialize $y_1 \in \mathbb{R}^{\mathcal{A}}$.
- For all $t = 1, \dots, T$:
  - Set $x^t = \Lambda(y_t)$.
  - Play $a_t \in \mathcal{A} \sim x^t \in \Delta(\mathcal{A})$ and receive c_{t,a^t}.
  - Set $\hat{c}_t = \frac{c_{t,a^t}}{x^t_{a_t}} e_{a_t}$.
  - Update $y_{t+1} = y_t - \gamma_t \hat{c}_t$

(EXP3) vs the Hedge algorithm

Hedge

- Require set of actions $\mathcal{A}$, sequence of cost vectors $c_t \in [0,1]^{\mathcal{A}}$
- Initialize $y_1 \in \mathbb{R}^{\mathcal{A}}$
- For all $t = 1, \dots, T$:
  - Set $x^t = \Lambda(y_t)$.
  - Play $x^t \in \Delta(\mathcal{A})$ and receive c_t.
  - Update $y_{t+1} = y_t - \gamma_t c_t$

EXP3

- Require set of actions $\mathcal{A}$, sequence of cost vectors $c_t \in [0,1]^{\mathcal{A}}$
- Initialize $y_1 \in \mathbb{R}^{\mathcal{A}}$
- For all $t = 1, \dots, T$:
  - Set $x^t = \Lambda(y_t)$.
  - Play $a_t \in \mathcal{A} \sim x^t \in \Delta(\mathcal{A})$ and receive c_{t,a^t}.
  - Set $\hat{c}_t = \frac{c_{t,a^t}}{x^t_{a_t}} e_{a_t}$.
  - Update $y_{t+1} = y_t - \gamma_t \hat{c}_t$

Remarks:

- Hedge and (EXP3) share a common algorithmic template.
- Crucial differences are as follows:
  - Hedge updates the whole cost vector (or a noise version of it) c_t.
  - (EXP3) updates the (IWE) estimators.

Regret analysis

Basic steps

- Consider constant step-size $\gamma \equiv \gamma_t$ (depending on the horizon T).
- Fix benchmark stategy $p \in \Delta(\mathcal{A})$ and define the respective Fenchel coupling:

$$F_t \equiv F(p, y_t) = \sum_{a \in \mathcal{A}} p_a \log p_a + \log \sum_{a \in \mathcal{A}} e^{y_{t,a}} - \langle y_t, p \rangle$$

A first natural approach

Use the generic (MD) energy inequality

Tentative proof

Main components

- Telescope and take expectations on both sides:

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle \hat{c}_t, x^t - p\rangle\right] \leq \frac{F_1}{\gamma} + \frac{\gamma}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\hat{c}_{a,t}\|^2\right]$$

- Apply unbiasedness and the upper for the second moment:

$$\mathbb{E}[Reg_p(T)] \leq \frac{F_1}{\gamma} + \frac{\gamma}{2}\mathcal{O}(\frac{1}{x_{a,t}})$$

Remarks:

- Observe that we have a (possibly) exploding term!
- How can we proceed?

A new energy inequality for (EXP3)

- As we observed the standard approach cannot be applied due to the irregular behaviour of the "error term".
- A new energy inequality is required!

Energy inequality for (EXP3)

Fix some $y \in \mathbb{R}^{\mathcal{A}}, w \in (-\infty, 1]^{\mathcal{A}}$ and let $x \propto e^{y}$. Then, the following inequality holds:

$$\log \sum_{a \in \mathcal{A}} e^{y_a - w_a} \leq \log \sum_{a \in \mathcal{A}} e^{y_a} + \langle x, w \rangle + \sum_{a \in \mathcal{A}} x_a w_a^2$$

- A main stepping stone is the following technical lemma:

Crucial inequality

For all $x \leq 1$, the following inequality holds:

$$e^x \leq 1 + x + x^2$$

Exercise: ○ Prove that the above inequality holds.

Energy inequality for (EXP3) (proof)

- We now proceed to the technical proof of the (EXP3) energy inequality:

Proof

We have:

$$\begin{aligned}\log \sum_{a\in\mathcal{A}} e^{y_a - w_a} &\leq \log \sum_{a\in\mathcal{A}} e^{y_a}(1 + w_a + w_a^2) = \log \sum_{a\in\mathcal{A}} e^{y_a} + \log \frac{\sum_{a\in\mathcal{A}} e^{y_a}(1 + w_a + w_a^2)}{\sum_{a\in\mathcal{A}} e^{y_a}} \\ &\leq \log \sum_{a\in\mathcal{A}} e^{y_a} + \log \sum_{a\in\mathcal{A}} x_a(1 + w_a + w_a^2) \\ &\leq \log \sum_{a\in\mathcal{A}} e^{y_a} + \sum_{a\in\mathcal{A}} x_a w_a + \sum_{a\in\mathcal{A}} x_a w_a^2\end{aligned}$$

Regret analysis continued

Proof for regret

- Use energy inequality, rearrange and telescope $t = 1, \dots T$:

$$\sum_{t=1}^{T} \langle \hat{c}_t, x_t - p \rangle \leq \frac{F_1}{\gamma} + \gamma \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} x_{a,t} \hat{c}_{\alpha,t}^2$$

- Take expectations on both sides:

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \hat{c}_t, x_t - p \rangle\right] \leq \frac{F_1}{\gamma} + \gamma \sum_{t=1}^{T} \mathbb{E}\left[\sum_{a \in \mathcal{A}} x_{a,t} \hat{c}_{\alpha,t}^2\right]$$

- To conclude, we need to deal with:

$$\mathbb{E}\left[\sum_{a \in \mathcal{A}} x_{a,t} \hat{c}_{\alpha,t}^2\right]$$

Bounding the residual

- We now proceed by bounding the new "error" term:

Proof of the second moment bound

$$\mathbb{E}\left[\sum_{a\in\mathcal{A}} x_{a,t}\hat{c}^2_{\alpha,t}\right] = \sum_{a'\in\mathcal{A}} x_{a',t} \sum_{a\in\mathcal{A}} \frac{\mathbf{1}_{a'=a}}{x^2_{a,t}} c^2_{a',t} = \sum_{a\in\mathcal{A}} x^2_{a,t} \frac{1}{x^2_{a,t}} c^2_{a,t}$$
$$= \sum_{a\in\mathcal{A}} c^2_{a,t} \leq |\mathcal{A}|$$

Hence, we have:

$$\sum_{t=1}^{T} \mathbb{E}\left[\sum_{a\in\mathcal{A}} x_{a,t}\hat{c}^2_{\alpha,t}\right] \leq |\mathcal{A}|T$$

Regret of (EXP3)

- Summarizing, we have the following generic regret bound:

Theorem (Regret for (EXP3))

Assume that x^t are the iterates generated by (EXP3). Then, the following inequality holds:

$$\mathbb{E}\left[\mathcal{R}_p(T)\right] \leq \frac{F_1}{\gamma} + \gamma |\mathcal{A}| T$$

Remark:

- For no-regret an appropriate (horizon-dependent) constant step-size needs to be selected!

Step-size selection

Step-size choice

We determine the step-size:

$$\mathbb{E}\left[\mathcal{R}_p(T)\right] \leq \underbrace{\frac{F_1}{\gamma} + \gamma |\mathcal{A}| T}_{\text{we minimize w.r.t. } \gamma}.$$

Compute the minimum

In general, we want to minimize a function of the following form:

$$f(\gamma) = \frac{a}{\gamma} + b\gamma \;\text{ with }\; \gamma > 0.$$

By finding the zeros of its first derivative, $f'(\gamma) = -\frac{a}{\gamma^2} + b = 0$ we have:

$$\gamma = \sqrt{\frac{2F_1}{T}}$$

No-regret for (EXP3)

Corollary (Guarantees of EXP3, Auer et al [4])

Assume that (EXP3) is run with a step-size $\gamma = \sqrt{\log|\mathcal{A}|/|\mathcal{A}|T}$. *Then , the following guarantee holds:*

$$\mathbb{E}\left[\mathcal{R}_p(T)\right] \leq 2\sqrt{|\mathcal{A}|\log|\mathcal{A}|T}.$$

Remarks:

- The above bound is tight in T, Abernethy et al [1].
- Worse than the full info bound by a factor of $\sqrt{|\mathcal{A}|}$.
- The regret can be improved to $\mathcal{O}(\sqrt{|\mathcal{A}|T})$ but no lower, Audibert & Bubeck [3].
- The step-size requires prior knowledge on the play horizon T.

Summary

- Establish the framework of online decision processes with bandit feedback.
- Define the appropriate oracle estimator tailored for the bandit feedback framework.
- Define the (EXP3) method.
- Define the Fenchel coupling and the appropriate energy inequalities.
- Establish regret guarantees for the said method.

References I

[1] Jacob D. Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari.
Optimal stragies and minimax lower bounds for online convex games.
In Rocco A. Servedio and Tong Zhang, editors, *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 415–424. Omnipress, 2008.
(Cited on page 29.)

[2] Kimon Antonakopoulos, E. Veronica Belmega, and Panayotis Mertikopoulos.
Online and stochastic optimization beyond lipschitz continuity: A riemannian approach.
In *International Conference on Learning Representations*, 2020.
(Cited on page 17.)

[3] Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi.
Regret in online combinatorial optimization, 2013.
(Cited on page 29.)

[4] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire.
Gambling in a rigged casino: The adversarial multi-armed bandit problem.
Levine's Working Paper Archive 462, David K. Levine, December 2010.
(Cited on page 29.)

References II

[5] Panayotis Mertikopoulos and Zhengyuan Zhou.
Learning in games with continuous action spaces and unknown payoff functions.
Mathematical Programming, Series A, 173(1-2):465–507, 2019.
(Cited on pages 16 and 17.)

[6] Yurii Nesterov.
Primal-dual subgradient methods for convex problems.
Mathematical programming, 120(1):221–259, 2009.
(Cited on page 15.)

[7] Shai Shalev-Shwartz.
Online learning and online convex optimization.
Foundations and Trends® in Machine Learning, 4(2):107–194, 2012.
(Cited on pages 14 and 15.)