CS272 - Theoretical Foundations of Learning, Decisions, and Games

Lecture 10: Bandits and Partial Feedback

September 30, 2025

Lecturer: Nika Haghtalab *Readings: not specified*
Scribe: Kazusato Oko

1 Follow the Perturbed Leader (continued)

In this lecture, despite the title, we mainly continue the analysis of the Follow the Perturbed Leader (FTPL) algorithm from the previous class. We begin by restating the FTPL algorithm, the theorem to be proved, and the lemma established in the previous lecture.

Algorithm 1 Follow the Perturbed Leader (FTPL)

input: total number of rounds T and perturbation parameter ϵ
for $t = 1, 2, \dots, T$ **do**
  sample $c^0 \sim \mathrm{Unif}([0, 2\epsilon^{-1}]^n)$
  $x^t = \arg\min_{x \in \mathcal{X}} \sum_{\tau=0}^{t-1} c^\tau \cdot x$
end for

For an adaptive adversary, it is important to re-draw c^0 at every step to prevent the adversary from inferring c^0 based on the algorithm's past actions. For an oblivious adversary, we could draw c^0 once before the for-loop and reuse it throughout, while still obtaining the same regret bound.

Theorem 1.1. *Assume that $\mathcal{X}$ is a domain of $\mathbb{R}^n$ such that $\max_{x,x' \in \mathcal{X}} \|x - x'\|_1 \le D$, and that the cost functions satisfy $\max_{1 \le t} \|c^t\|_1 \le 1$, and that c^0 is sampled from $\mathrm{Unif}([0, 2\epsilon^{-1}]^n)$. Then, Algorithm 1 (FTPL) achieves*

$$\mathbb{E}[\textit{Regret}] \le \underbrace{\frac{2D}{\epsilon}}_{①} + \underbrace{\frac{T\epsilon D}{2}}_{②}. \tag{1}$$

By taking $\epsilon = \frac{2}{\sqrt{T}}$, we have $\mathbb{E}[\textit{Regret}] \le 2D\sqrt{T}$.

Lemma 1.2. *Algorithm 1 (FTPL) gives us the following guarantee for any $x \in \mathcal{X}$, the regret with respect to x is*

$$\textit{Regret}(x) := \sum_{t=1}^{T} c^t \cdot x^t - \sum_{t=1}^{T} c^t \cdot x \le \underbrace{c^0 \cdot (x - x^1)}_{①} + \underbrace{\sum_{t=1}^{T} c^t \cdot (x^t - x^{t+1})}_{②}. \tag{2}$$

Eq. (1) can be decomposed into ① the perturbation term and ② the stability term. The former corresponds to the degradation in prediction caused by c^0, while the latter corresponds to the reduction of $x^t - x^{t+1}$ due to the perturbation. The proof proceeds by evaluating the two terms on (2) of Lemma 1.2. The terms ① and ② in (2) correspond respectively to the two terms in (1).

Proof of Theorem 1.1. We first evaluate ①. From $\max_{x,x' \in \mathcal{X}} \|x - x'\|_1 \leq D$. Moreover, $c^0 \sim \mathrm{Unif}([0, 2\epsilon^{-1}]^n)$, which implies $\|c^0\|_\infty \leq 2\epsilon^{-1}$, we obtain

$$(\text{① of (2)}) = c^0 \cdot (x - x^0) \leq \|c^0\|_\infty \|x - x^0\|_1 \leq \frac{2D}{\epsilon}. \tag{3}$$

On the other hand, for ②, it suffices to prove $c^t \cdot \mathbb{E}[(x^t - x^{t+1})] \leq \frac{\epsilon D}{2}$ for all $t = 1, 2, \ldots, T$. To this end, let us recall the definitions of x^t and x^{t+1}:

$$x^t = \arg\min_{x \in \mathcal{X}} \sum_{\tau=0}^{t-1} c^\tau \cdot x = \arg\min_{x \in \mathcal{X}} \left[\left(\sum_{\tau=1}^{t-1} c^\tau + c^0 \right) \cdot x \right],$$
$$x^{t+1} = \arg\min_{x \in \mathcal{X}} \sum_{\tau=0}^{t} c^\tau \cdot x = \arg\min_{x \in \mathcal{X}} \left[\left(\sum_{\tau=1}^{t-1} c^\tau + (c^0 + c^t) \right) \cdot x \right].$$

Here, observe that the distributions of c^0 and $c^0 + c^t$ have significant overlap. As a result, x^t when $c^0 = c$ coincides with x^{t+1} when $c^0 + c^t = c$, which allows us to cancel out most of $x^t - x^{t+1}$.

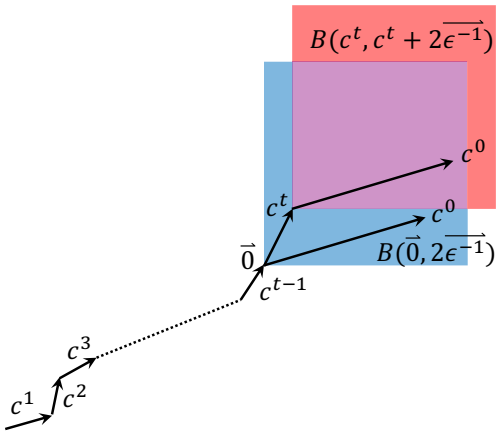

Figure 1: An illustration of the vector decomposition of the cost. In particular, when we re-center at $\sum_{\tau=1}^{t-1} c^\tau$, the distributions of c^0 and $c^0 + c^t$ become the uniform distributions over $B(\vec{0}, 2\overrightarrow{\epsilon^{-1}})$ and $B(c^t, c^t + 2\overrightarrow{\epsilon^{-1}})$, respectively, and these distributions largely overlap.

To formalize this, for $a, b \in \mathbb{R}^n$ we introduce $B(a, b) := \{ x \in \mathbb{R}^n \mid a_i \leq x_i \leq b_i \ \ (i = 1, \ldots, n) \}$. Moreover, for a scalar p, we use $\vec{p}$ to denote the vector $(p, \ldots, p)^\top \in \mathbb{R}^n$. We denote $B(\vec{0}, 2\overrightarrow{\epsilon^{-1}}) \setminus B(c^t, c^t + 2\overrightarrow{\epsilon^{-1}})$ by the blue, $B(c^t, c^t + 2\overrightarrow{\epsilon^{-1}}) \setminus B(\vec{0}, 2\overrightarrow{\epsilon^{-1}})$ by the red, and $B(\vec{0}, 2\overrightarrow{\epsilon^{-1}}) \cup B(c^t, c^t + 2\overrightarrow{\epsilon^{-1}})$ by the purple. Looking at Figure 1, c^0 follows the uniform distribution over

$B(\vec{0}, 2\overrightarrow{\epsilon^{-1}})$ (blue + purple), while $c^0 + c^t$ follows the uniform distribution over $B(c^t, c^t + 2\overrightarrow{\epsilon^{-1}})$ (red + purple). Then,

$$\mathbb{E}[x^t] = \Pr[c^0 \in \text{purple}]\ \mathbb{E}[x^t \mid c^0 \in \text{purple}] + \Pr[c^0 \in \text{blue}]\ \mathbb{E}[x^t \mid c^0 \in \text{blue}] \tag{4}$$

$$\begin{aligned}\mathbb{E}[x^{t+1}] = {} & \Pr[c^0 + c^t \in \text{purple}]\ \mathbb{E}[x^{t+1} \mid c^0 + c^t \in \text{purple}] \\ & + \Pr[c^0 + c^t \in \text{red}]\ \mathbb{E}[x^{t+1} \mid c^0 + c^t \in \text{red}]\end{aligned} \tag{5}$$

Note that the probability density is identical for both distributions at every point in their support. Combining this with the fact that x^t when $c^0 = c$ is equal to x^{t+1} when $c^0 + c^t = c$, we obtain $\Pr[c^0 \in \text{purple}] = \Pr[c^0 + c^t \in \text{purple}]$, and $\mathbb{E}[x^t \mid c^0 \in \text{purple}] = \mathbb{E}[x^{t+1} \mid c^0 + c^t \in \text{purple}]$. Therefore,

$$(4) - (5) = \Pr[c^0 \in \text{blue}]\ \mathbb{E}[x^t \mid c^0 \in \text{blue}] - \Pr[c^0 + c^t \in \text{red}]\ \mathbb{E}[x^{t+1} \mid c^0 + c^t \in \text{red}] \tag{6}$$

Here, we use the fact that the blue and purple regions, or the non-overlapping regions, are relatively small. Because of the symmetry, we have $\Pr[c^0 \in \text{blue}] = \Pr[c^0 + c^t \in \text{red}]$, so we focus on the blue region. That is decomposed as

$$(\text{blue region}) = B(\vec{0}, 2\overrightarrow{\epsilon^{-1}}) \setminus B(c^t, c^t + 2\overrightarrow{\epsilon^{-1}}) = \bigcup_{i=1}^{n} (B(\vec{0}, 2\overrightarrow{\epsilon^{-1}}) \cup \{x \in \mathbb{R}^n \mid 0 \le x_i \le c_i^t\}).$$

Each set $B(\vec{0}, 2\overrightarrow{\epsilon^{-1}}) \cup \{x \in \mathbb{R}^n \mid 0 \le x_i \le c_i^t\}$ occupies a fraction of $c_i^t \times \frac{\epsilon}{2}$ of the entire $B(\vec{0}, 2\overrightarrow{\epsilon^{-1}})$. Therefore, comparing the volumes,

$$\frac{(\text{blue region})}{B(\vec{0}, 2\overrightarrow{\epsilon^{-1}})} \le \sum_{i=1}^{n} c_i^t \times \frac{\epsilon}{2} = \|c^t\|_1 \frac{\epsilon}{2} \le \frac{\epsilon}{2},$$

where we applied the assumption that $\|c^t\|_1 \le 1$. The LHS is equal to $\Pr[c^0 \in \text{blue}]$, so we have that

$$\Pr[c^0 \in \text{blue}] = \Pr[c^0 + c^t \in \text{red}] \le \frac{\epsilon}{2}.$$

Combining this with $\|c^t\|_\infty \le 1$ and $\max_{x,x' \in \mathcal{X}} \|x - x'\|_1 \le D$, (6) is further bounded by

$$c^0 \cdot ((4) - (5)) = \frac{\epsilon}{2} \|c^t\|_\infty \max_{x,x' \in \mathcal{X}} \|x - x'\|_1 \le \frac{\epsilon D}{2}.$$

Therefore,

$$(② \text{ of } (2)) \le (T-1)\frac{\epsilon D}{2} \le \frac{T \epsilon D}{2}. \tag{7}$$

Now, (3) and (7) bound the regret to any x, and thus we obtain the desired bound (1). □

Historical remarks. The FTPL algorithm in the online learning setting and its guarantee for linear costs were introduced by Kalai and Vempala [2005]. In general, algorithmic gaps exist between offline and online learning; for example, there are classes where ERM is efficient but no oracle-efficient online algorithm exists [Hazan and Koren, 2016]. However, when smoothness constraints are imposed on the adaptive adversary, an FTPL-based online algorithm achieves no-regret guarantees with polynomial-time complexity [Haghtalab et al., 2024].

2 Partial Feedback

In the setting we have studied so far, at each time t we can observe the costs c_i^t of all experts. An example of this is classification, where given (x^t, y^t) we can check for each $h_i(x^t)$ whether it matches y^t.

On the other hand, there are cases where not all c_i^t are observable.

- **Online routing.** We aim to find the shortest travel time from point A to point B. Here, the travel time of each path is determined by traffic. However, the traffic is only revealed for the path that is actually selected and traversed.
- **Pricing.** We aim to set a price for an item and sell it at the highest possible price. For example, suppose we set the price to \$3 and the item is sold. However, we do not know whether it would have sold at \$10 as well.

From now on, we consider online learning with partial information, as described above. Formally, suppose there is a known family of experts $E_1, \ldots, E_m$, and when we choose $j \in [m]$, the observable costs are $\{c_i^t \mid i \in E_j\}$. The case $m = 1,\ E_1 = [n]$ corresponds to full information (our previous setting), while the case $m = n,\ E_i = \{i\}$ corresponds to multi-armed bandits. For the lectures, we will focus on the multi-armed bandit setting. Though, this abstraction is more general and allows us to study other intermediate settings as well (e.g., in the homeworks and optional readings).

3 Multi-Armed Bandits

From here, we turn to the setting of multi-armed bandits. In the full-information case, the regret was $\sqrt{T \log n}$, but how does the dependence on n and T change in the case of multi-armed bandits? When designing efficient algorithms, two key considerations arise:

- **How to choose an expert?** That is, we must balance exploring new possibilities (exploration) with exploiting those that have proven useful in the past (exploitation).
- **How to update an expert's weight?** Since different experts are selected with different probabilities, the penalization when updating the weights has to take into account the likelihood of being picked.

At the end of the lecture, we introduce the "old school" Exponential Weights for Exploration and Exploitation (EXP3) algorithm [Auer et al., 2002]. The proof will be deferred to next week.

Algorithm 2 Exponential Weights for Exploration and Exploitation (EXP3)

input: total number of rounds T and exploration parameter γ
initialize: weights $w_i^1 = 1$ for all $i = 1, \ldots, n$
for $t = 1, 2, \ldots, T$ **do**
  define probability distribution $q_i^t = (1-\gamma)\frac{w_i^t}{\sum_{j=1}^n w_j^t} + \frac{\gamma}{n}$ for all j
  sample expert $i^t \sim q^t$ and observe cost $c_{i^t}^t$
  update weight $w_j^{t+1} = w_j^t \cdot \exp\big(\gamma \hat{c}_j^t / n\big)$ for all j, where $\hat{c}_j^t = 0$ for all $j \neq i^t$, with $\hat{c}_{i^t}^t = \frac{c_{i^t}^t}{q_{i^t}^t}$.
end for

References

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis with adaptive adversaries. *Journal of the ACM*, 71(3):1–34, 2024.

Elad Hazan and Tomer Koren. The computational power of optimization in online learning. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 128–141, 2016.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.