Online and Reinforcement Learning (2025)
Home Assignment 1

Your name and student ID

Contents

1 Find an online learning problem from real life

Exercise 5.1 (Find an online learning problem from real life). Find two examples of real life problems that fit into the online learning framework (online, not reinforcement!). For each of the two examples explain what is the set of actions an algorithm can take, what are the losses (or rewards) and what is the range of the losses/rewards, whether the problem is stateless or contextual, and whether the problem is i.i.d. or adversarial, and with full information or bandit feedback

1.1 Post Recommendation on Social Media

Deciding in real time which article or advertisement to display to a user.

- **Actions:**
  The algorithm chooses one item (news article or ad) from a finite set of options.
- **Reward:**
  1 if the user clicks on the suggested content, 0 if the user does not (range [0, 1]).
  In more sophisticated systems, the reward could be based on the time spent by the user on the suggested content.
- **Stateless vs. Contextual:**
  In the simplest implementation, it could be stateless, basing recommendations exclusively on the outcomes of previous interactions. However, in actual implementations, it is contextual, as the algorithm considers the user's profile and related information.

- **Environment (i.i.d. vs. Adversarial):**
  While an idealized model might assume users arrive from an i.i.d. process, real-world user behavior is often adversarial (or non-stationary) as preferences and trends shift over time.

- **Feedback:**
  The feedback is bandit feedback; the algorithm only observes the outcome (click or no click) for the displayed item, not for all items in the set.

1.2 Pricing of Products Over Time

Deciding in real time with what price to sell a product to maximize profit.

- **Set of Actions:**
  Each day, the algorithm can change the price of a product, selecting from a fixed number of price options.

- **Reward:**
  The reward is defined as the profit obtained at the end of the day, which is the revenue from sales minus the cost of production.
  The range of the reward is $(0, +\infty)$ since we assume that none of the prices are lower than the production cost.

- **Stateless vs. Contextual:**
  The problem could be considered stateless if the algorithm only considers the prices and profits of the previous days.

- **Environment (i.i.d. vs. Adversarial):**
  The environment could be considered i.i.d. if the demand for the product remains constant over time. However, in reality, it is adversarial, as demand can change over time, especially for products that are not purchased repeatedly.

- **Feedback Type:**
  The feedback in dynamic pricing is of the bandit type, as the algorithm only observes the reward for the chosen action (applied price).

2 Follow The Leader (FTL) algorithm for i.i.d. full information games

Setting

- We have **two actions**, a^* and a, with rewards in $[0, 1]$.

- The rewards for each action are *i.i.d. across rounds*, and *fully observed* each round.

- Let $\mu(a)$ be the true expected reward of action a. Without loss of generality, let

$$a^* \;=\; \arg\max_a \mu(a).$$

- Define the *gap* of the suboptimal arm by

$$\Delta \;=\; \mu(a^*) - \mu(a) \;>\; 0.$$

- We consider T rounds, indexed by $t = 1, 2, \ldots, T$.

Algorithm: Follow the Leader (FTL)

- **Initialization:** In the very first round(s), you may pick any arm(s).
- **At each round $t \geq 2$:**
  1. For each arm $b \in \{a, a^*\}$, compute its *empirical mean* based on *all* observed rewards of b so far:
  $$\hat{\mu}_{t-1}(b) \;=\; \frac{1}{t-1} \sum_{s=1}^{t-1} X_s(b),$$
  where $X_s(b)$ is the reward observed for arm b at round s.
  2. *Choose* the arm whose empirical mean is larger:
  $$A_t \;=\; \arg\max_{b \in \{a, a^*\}} \hat{\mu}_{t-1}(b).$$

Because this is a **full-information** setting, we observe $X_t(a)$ and $X_t(a^*)$ for both arms at every round t, not just the chosen one.

Key Idea: Probability of Confusion

Let us define the event

$$\big\{\text{“FTL picks the suboptimal arm } a \text{ at round } t\text{”}\big\} \quad\Longleftrightarrow\quad \big\{\,\hat{\mu}_{t-1}(a) \;\geq\; \hat{\mu}_{t-1}(a^*)\big\}.$$

Because $\Delta = \mu(a^*) - \mu(a) > 0$, the only way for $\hat{\mu}_{t-1}(a)$ to overtake $\hat{\mu}_{t-1}(a^*)$ is if

$$\big[\hat{\mu}_{t-1}(a) - \mu(a)\big] \;-\; \big[\hat{\mu}_{t-1}(a^*) - \mu(a^*)\big] \;\geq\; \Delta.$$

By Hoeffding's inequality (for $[0,1]$-bounded i.i.d. samples),

$$\mathbb{P}\Big(\hat{\mu}_n(b) - \mu(b) \;\geq\; \epsilon\Big) \;\leq\; \exp\big(-2\,n\,\epsilon^2\big),$$

one sees that

$$\mathbb{P}\Big(\hat{\mu}_{t-1}(a) \;\geq\; \hat{\mu}_{t-1}(a^*)\Big) \;\leq\; \exp\big(-c\,(t-1)\,\Delta^2\big) \quad \text{for some constant } c > 0.$$

We will sometimes refer to this as the *probability of confusion* at round t.

Bounding the Regret

Regret definition. The (pseudo-)regret is

$$R_T \;=\; \sum_{t=1}^{T} \big[\mu(a^*) \;-\; \mu(A_t)\big].$$

Since $A_t \in \{a, a^*\}$, the only time we incur regret $\Delta > 0$ is precisely when $A_t = a$. Thus,

$$R_T \;=\; \Delta \sum_{t=1}^{T} \mathbf{1}\{\text{FTL picks } a \text{ at round } t\}.$$

Taking expectation,

$$\mathbb{E}[R_T] \;=\; \Delta \sum_{t=1}^{T} \mathbb{P}\big(A_t = a\big) \;=\; \Delta \sum_{t=1}^{T} \mathbb{P}\Big(\hat{\mu}_{t-1}(a) \;\geq\; \hat{\mu}_{t-1}(a^*)\Big).$$

Using the exponential bound from above, define

$$\delta(t) \;=\; \exp\big(-c\,(t-1)\,\Delta^2\big).$$

Hence,

$$\mathbb{E}[R_T] \;\leq\; \Delta \sum_{t=1}^{T} \delta(t).$$

Geometric series. Note that $\sum_{t=1}^{\infty} \exp(-c\,(t-1)\,\Delta^2)$ converges (it is a geometric series). Concretely,

$$\sum_{t=1}^{\infty} \exp\big(-c\,(t-1)\,\Delta^2\big) \;=\; \frac{1}{1 - e^{-\,c\,\Delta^2}} \;<\; \infty.$$

Hence there is a constant $C(\Delta)$ such that

$$\sum_{t=1}^{T} \exp\big(-c\,(t-1)\,\Delta^2\big) \;\leq\; C(\Delta), \quad \text{independently of } T.$$

Multiplying by Δ,

$$\mathbb{E}[R_T] \;\leq\; \Delta\, C(\Delta).$$

That is, *the total expected regret* remains **bounded** by a constant with respect to T.

Final Statement of the Bound

Because FTL eventually "locks onto" the better arm (and stays there with high probability), we conclude

$$R_T \;=\; O(1) \quad \text{as } T \to \infty.$$

A more explicit expression for the constant bound is

$$\mathbb{E}[\,R_T\,] \;\le\; \Delta\,\frac{1}{1-e^{-\,c\,\Delta^2}} \;=\; \frac{\Delta}{1-e^{-\,c\,\Delta^2}},$$

which is finite for every $\Delta > 0$. If there are multiple suboptimal arms $a = 1, \ldots, K-1$, one simply sums similar terms for each suboptimal $\Delta(a)$; the result is still a finite constant.

Comparison with the Bandit Setting

In the professor's *bandit* slides, one observes only the reward of the chosen arm each round and thus must *explore* explicitly, leading to a regret growing like $\log T$. Here, in *full-information* feedback, both arms' rewards are seen every time, so $\hat{\mu}_t(a)$ and $\hat{\mu}_t(a^*)$ concentrate exponentially fast for *both* arms. As a result, FTL typically makes only a finite number of mistakes (suboptimal plays), yielding a **constant** regret bound.