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

Exercise 5.2 (Follow The Leader (FTL) algorithm for i.i.d. full information games). Follow the leader (FTL) is a playing strategy that at round t plays the action that was most successful up to round t ("the leader"). Derive a bound for the pseudo regret of FTL in i.i.d. full information games with K possible actions and outcomes bounded in the [0, 1] interval (you can work with rewards or losses, as you like).

1. Algorithm (FTL)

- **Initialization:** At round $t = 1$, pick any action (or pick each action once if you prefer to break ties).
- **For each round $t \geq 2$:**
  1. Compute the empirical average reward $\hat{\mu}_{t-1}(a)$ of each action a based on *all past observations* of that action:
  $$\hat{\mu}_{t-1}(a) \;=\; \frac{1}{t-1}\sum_{s=1}^{t-1} X_s(a),$$
  where $X_s(a)$ is the reward of action a at round s.
  2. Play the action
  $$A_t \;=\; \arg\max_a \; \hat{\mu}_{t-1}(a).$$
  3. Break ties arbitrarily if needed.

Because we are in a *full-information* setting, each round's reward for *all* actions is observed, not only the one played.

2. Notation and Goal

- Let $\mu(a)$ be the true expected reward of action a.
- Let a^* be an optimal action, *i.e.*,
$$\mu(a^*) \;=\; \max_a \mu(a).$$
- Define the *gap* for a suboptimal action $a \neq a^*$ by
$$\Delta(a) \;=\; \mu(a^*) \;-\; \mu(a) \;>\; 0.$$
- The *pseudo-regret* over T rounds is
$$R_T \;=\; \sum_{t=1}^{T}\Big[\mu(a^*) - \mu(A_t)\Big].$$

  We aim to bound R_T.

3. Key Event: Picking a Suboptimal Action

For a single suboptimal action $a \neq a^*$, FTL picks a at round t if and only if

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

Because the rewards are i.i.d. *and* we have full information, each $\hat{\mu}_{t-1}(a)$ is the average of $(t-1)$ i.i.d. samples with mean $\mu(a)$. Subtracting $\hat{\mu}_{t-1}(a^*)$ from $\hat{\mu}_{t-1}(a)$ yields:

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

Hence the event $\{\hat{\mu}_{t-1}(a) \geq \hat{\mu}_{t-1}(a^*)\}$ implies

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

This sort of "overtaking" event becomes very unlikely once t grows, by standard concentration inequalities.

4. Bounding the Probability of Overtaking

Using Hoeffding's (or Chernoff) bound for $[0,1]$-valued i.i.d. rewards, we have: if $\hat{\mu}_n(a)$ is the empirical average of n i.i.d. samples with mean $\mu(a)$, then for any $\epsilon > 0$,

$$\mathbb{P}\bigl(\hat{\mu}_n(a) - \mu(a) \;\geq\; \epsilon\bigr) \;\;\leq\;\; \exp\bigl(-2\,n\,\epsilon^2\bigr).$$

Applying this to both $\hat{\mu}_{t-1}(a)$ and $\hat{\mu}_{t-1}(a^*)$ leads to a bound of the form

$$\mathbb{P}\Bigl[\hat{\mu}_{t-1}(a) \;\geq\; \hat{\mu}_{t-1}(a^*)\Bigr] \;\;\leq\;\; \exp\bigl(-c\,(t-1)\,\Delta(a)^2\bigr)$$

for some positive constant c. Because these probabilities decay exponentially in t, the expected number of times a suboptimal arm "overtakes" the optimal one is finite.

5. Summing Over t: Finite Number of Mistakes

Summing $\exp(-c(t-1)\Delta(a)^2)$ from $t = 1$ to ∞ gives a convergent geometric series. Hence, *in expectation*, each suboptimal arm $a \neq a^*$ is chosen only a finite (constant) number of times (with respect to T). That is, for each $a \neq a^*$,

$$\mathbb{E}[\,N_T(a)\,] \;\leq\; \text{(some constant depending on } \Delta(a) \text{ but not on } T).$$

Since each time we pick $a \neq a^*$ we suffer regret $\Delta(a)$, the total regret from arm a is at most

$$\Delta(a)\,\mathbb{E}[\,N_T(a)\,],$$

which remains constant in T. Summing over all suboptimal arms concludes the argument.

6. Final Regret Bound

Overall, because each suboptimal arm is played only a constant (in T) number of times in expectation,

$$R_T \;=\; \sum_{t=1}^{T} [\mu(a^*) - \mu(A_t)] \;\leq\; \sum_{a\colon \Delta(a)>0} \bigl[\text{constant depending on } \Delta(a)\bigr].$$

Hence R_T is $O(1)$ in T. Concretely, a typical form is

$$R_T \;\leq\; \sum_{a\colon \Delta(a)>0} \left(\frac{\text{constant}}{\Delta(a)^2}\right) \Delta(a) \;=\; \sum_{a\colon \Delta(a)>0} (\text{constant factor})\,,$$

confirming that *the regret does not grow with T.*

7. Comparison with Bandits

In the *full-information* case, we observe rewards of all actions each round, leading to fast (exponential) concentration of their empirical means. Thus, FTL effectively “locks onto” the optimal action and stops making mistakes.

In contrast, in the *bandit* setting, one only observes the reward of the *played* arm, which forces explicit exploration. The regret then typically grows *logarithmically* with T, rather than staying constant.

Summary: In the i.i.d. full-information setting, Follow the Leader has a *constant* (in T) regret bound, unlike the bandit setting where regret grows at least on the order of $\log T$.