Adversarial Multi-Armed Bandits

Chen-Yu Wei

Adversarial Multi-Armed Bandits

Given: set of arms $\mathcal{A} = \{1, \dots, A\}$

For time $t = 1, 2, \dots, T$:

Environment decides the reward vector $r_t = (r_t(1), \dots, r_t(A))$ (not revealing)

Learner chooses an arm $a_t \in \mathcal{A}$

Learner observes $r_t(a_t)$

$$\text{Regret} = \max_{a \in \mathcal{A}} \sum_{t=1}^{T} r_t(a) - \sum_{t=1}^{T} r_t(a_t)$$

Exponential Weight Updates for Bandits

$$p_{t+1}(a) = \frac{p_t(a)\exp(\eta r_t(a))}{\sum_{a' \in \mathcal{A}} p_t(a')\exp(\eta r_t(a'))}$$

Exponential Weight Updates for Bandits

No longer observable

$$p_{t+1}(a) = \frac{p_t(a)\exp(\eta r_t(a))}{\sum_{a' \in \mathcal{A}} p_t(a')\exp(\eta r_t(a'))}$$

- Only update the arm that we choose?

Exponential Weight Updates for Bandits

$$p_{t+1}(a) = \frac{p_t(a)\exp(\eta\hat{r}_t(a))}{\sum_{a'\in\mathcal{A}} p_t(a')\exp(\eta\hat{r}_t(a'))}$$

- $\hat{r}_t(a)$ is an **“estimator”** for $r_t(a)$
- But we can only observe the reward of one arm!
- Furthermore, $r_t(a)$ is different in every round (If I did not sample arm a in round t, I’ll never be able to estimate $r_t(a)$ in the future)

Unbiased Reward / Gradient Estimator

$$\mathcal{H}_t = \left(a_1, r_1(a_1), \cdots, a_{t-1}, r_{t-1}(a_{t-1})\right)$$

Inverse Propensity Weighting

$$\hat{r}_t(a) = \frac{r_t(a)}{p_t(a)}\mathbb{I}\{a_t = a\} = \begin{cases} \dfrac{r_t(a)}{p_t(a)} & \text{if } a_t = a \\ 0 & \text{otherwise} \end{cases}$$

$$\forall a, \quad \mathbb{E}\left[\hat{r}_t(a) \middle| \mathcal{H}_t\right] = \mathbb{E}\left[\frac{r_t(a)}{p_t(a)}\mathbb{I}\{a_t = a\} \middle| \mathcal{H}_t\right] = \frac{r_t(a)}{p_t(a)} \underbrace{\mathbb{E}\left[\mathbb{I}(a_t = a) \middle| \mathcal{H}_t\right]}_{p_t(a)}$$

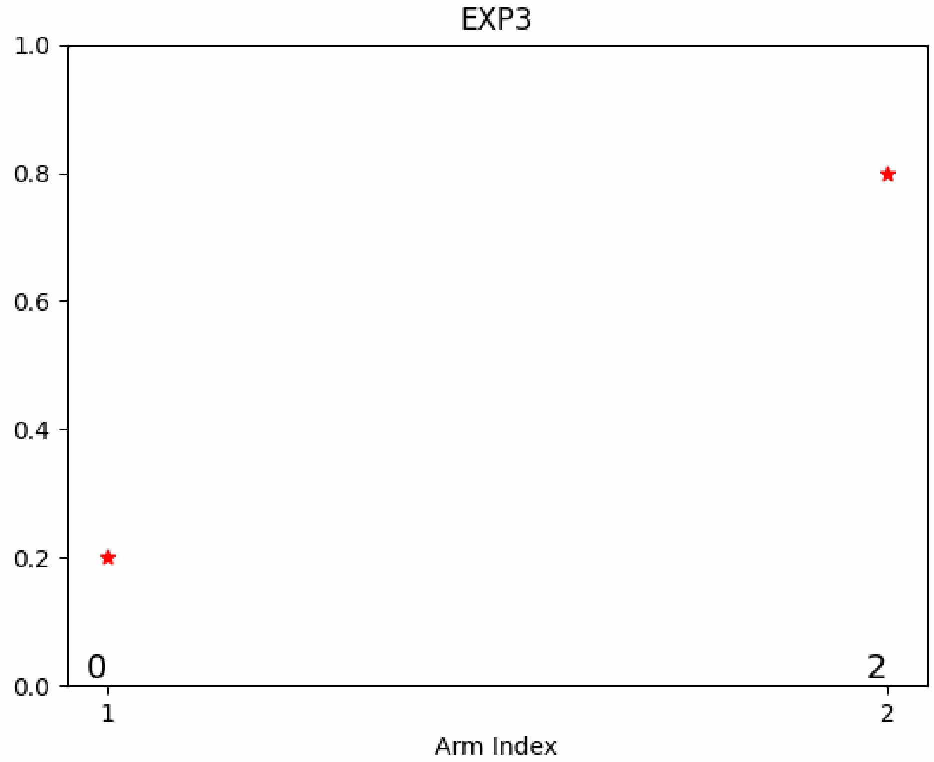
$$= r_t(a)$$

Directly Applying Exponential Weights

```
$p_1(a) = 1/A$ for all $a$
For $t = 1,2,\dots,T$:
    Sample $a_t$ from $p_t$, and observe $r_t(a_t)$
    Define for all $a$:
        $\hat{r}_t(a) = \frac{r_t(a)}{p_t(a)} \mathbb{I}\{a_t = a\}$
    Update policy:
        $p_{t+1}(a) = \frac{p_t(a) \exp(\eta \hat{r}_t(a))}{\sum_{a' \in \mathcal{A}} p_t(a') \exp(\eta \hat{r}_t(a'))}$
```

Simple Experiment

- $A = 2,\ T = 1500,\ \eta = 1/\sqrt{T}$
- For $t \leq 500,\ r_t = [\text{Bernoulli}(0.2), \text{Bernoulli}(0.8)]$
- For $500 < t \leq 1500,\ r_t = [\text{Bernoulli}(0.8), \text{Bernoulli}(0.2)]$



Applying the Theorem

Theorem.

Assume that $\eta \hat{r}_t(a) \leq 1$ for all t, a. Then EWU

$$p_{t+1}(a) = \frac{p_t(a)\exp(\eta \hat{r}_t(a))}{\sum_{a' \in \mathcal{A}} p_t(a')\exp(\eta \hat{r}_t(a'))}$$

ensures for any $a^\star$,

$$\sum_{t=1}^{T} (\hat{r}_t(a^\star) - \langle p_t, \hat{r}_t \rangle) \leq \frac{\ln A}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{A} p_t(a) \hat{r}_t(a)^2$$

Several Issues / Questions

- The assumption $\eta \hat{r}_t(a) \leq 1$ may not be satisfied
- How are the left-hand side and the regret definition related?

$$\sum_{t=1}^{T} \left(\hat{r}_t(a^\star) - \langle p_t, \hat{r}_t \rangle\right) \quad \text{vs.} \quad \sum_{t=1}^{T} \left(r_t(a^\star) - r_t(a_t)\right)$$

- How to bound the term on the right hand side?

$$\eta \sum_{t=1}^{T} \sum_{a=1}^{A} p_t(a) \hat{r}_t(a)^2$$

How is the LHS related to the Regret?

$$\mathbb{E}\left[\sum_t \hat{r}_t(a^*) - \sum_t \underbrace{\langle p_t, \hat{r}_t \rangle}\right] = \mathbb{E}\left[\sum_t r_t(a^*)\right] - \mathbb{E}\left[\sum_t r_t(a_t)\right]$$

$$\downarrow$$

$$\sum_a \cancel{p_t(a)} \cdot \frac{r_t(a)}{\cancel{p_t(a)}} \mathbb{I}\{a_t = a\}$$

$$= r_t(a_t)$$

$$\sum_t \mathbb{E}\left[\langle p_t, \hat{r}_t \rangle\right]$$

$$= \sum_t \sum_a p_t(a)\, \mathbb{E}\left[\hat{r}_t(a)\right]$$

$$= \sum_t \sum_a p_t(a)\, r_t(a) = \sum_t \langle p_t, r_t \rangle$$

How to bound the term on the right-hand side?

$$\sum_a p_t(a)\,\hat{r}_t(a)^2 = \sum_a p_t(a)\cdot\left(\frac{r_t(a)}{p_t(a)}\mathbb{1}\{a_t=a\}\right)^2$$

$$= \sum_a p_t(a)\cdot\frac{r_t(a)^2}{p_t(a)^2}\mathbb{1}\{a_t=a\}$$

$$= \sum_a \frac{\mathbb{1}\{a_t=a\}}{p_t(a)}\boxed{r_t(a)^2} \le \sum_a \frac{\mathbb{1}\{a_t=a\}}{p_t(a)}$$

$$\mathbb{E}\left[\quad \cdot\cdot \quad\right]$$

$$\le \sum_a \mathbb{E}\left[\frac{\mathbb{1}\{a_t=a\}}{p_t(a)}\right] \le A$$

The assumption $\eta\hat{r}_t(a) \le 1$ is not satisfied

$$\mathbb{E}\left(\eta \cdot \frac{r_t(a)}{p_t(a)} \underline{\underline{\mathbb{1}\{a_t=a\}}}\right) = \eta\, r_t(a) \le 1$$

Solution 1: Adding Extra Exploration

- **Idea:** use at least η probability to choose each arm

- Instead of sampling a_t according to p_t, use

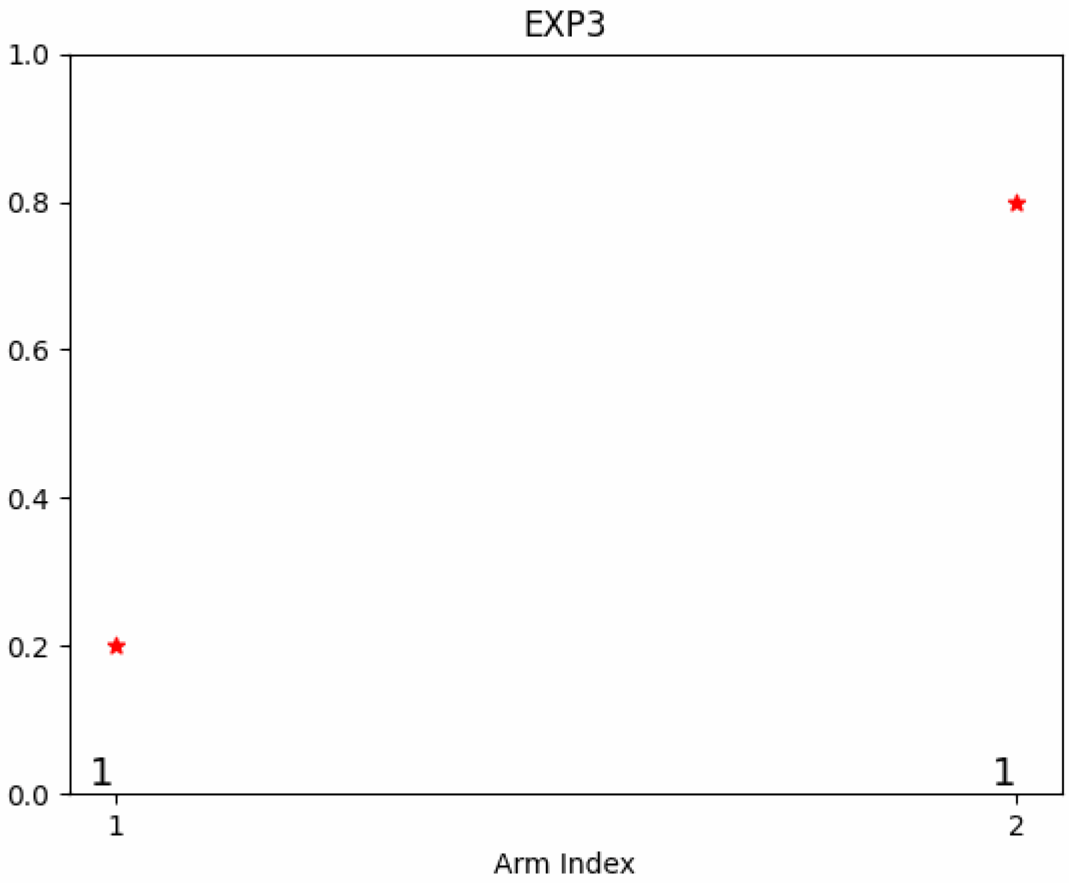
$$p_t'(a) = (1 - A\eta)p_t(a) + \eta$$

$$p_t' = (1-A\eta)\,p_t + A\eta \cdot \text{uniform}$$

Then the unbiased reward estimator becomes

$$\hat{r}_t(a) = \frac{r_t(a)}{p_t'(a)}\mathbb{I}\{a_t = a\} = \frac{r_t(a)}{(1 - A\eta)p_t(a) + \eta}\mathbb{I}\{a_t = a\} \le 1$$

Solution 1: Adding Extra Exploration



Applying Solution 1

```
$p_1(a) = 1/A$  for all $a$
For $t = 1,2, \ldots, T$:
    Sample $a_t$ from $p'_t = (1 - A\eta)p_t + A\eta \ \mathrm{uniform}(\mathcal{A})$, and observe $r_t(a_t)$
    Define for all $a$:
        $\hat{r}_t(a) = \frac{r_t(a)}{p'_t(a)} \mathbb{I}\{a_t = a\}$
    Update policy:
        $p_{t+1}(a) = \frac{p_t(a) \exp(\eta \hat{r}_t(a))}{\sum_{a' \in \mathcal{A}} p_t(a') \exp(\eta \hat{r}_t(a'))}$
```

Solution 2: Construct a Different Reward Estimator

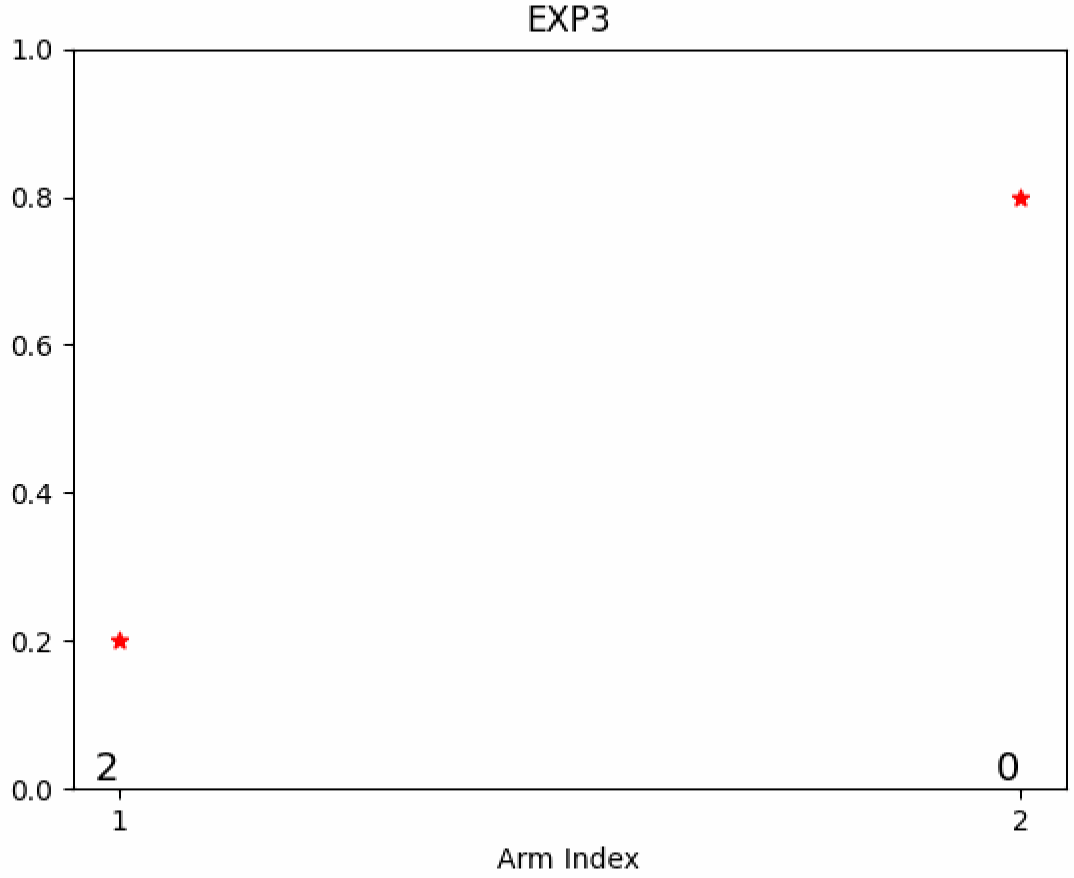
- Notice that the condition is only $\eta\hat{r}_t(a) \leq 1$. The reward estimator is allowed to be **very negative**! (Check our proof)

- Still sample a_t from p_t, but construct the reward estimator as

$$\hat{r}_t(a) = \frac{r_t(a) - 1}{p_t(a)} \mathbb{I}\{a_t = a\} + 1$$

- Why this resolves the issue?

Solution 2: Construct a Different Reward Estimator


EXP3
Arm Index