Adversarial Multi-Armed Bandits

Adversarial Multi-Armed Bandits

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

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

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

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

Learner observes $r_t(a_t)$

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

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

Inverse Propensity Weighting

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

Directly Applying Exponential Weights

$p_1(a) = 1/A$ for all a

For $t = 1,2,\ldots,T$:

Sample a_t from p_t, and observe $r_t(a_t)$

Define for all a:

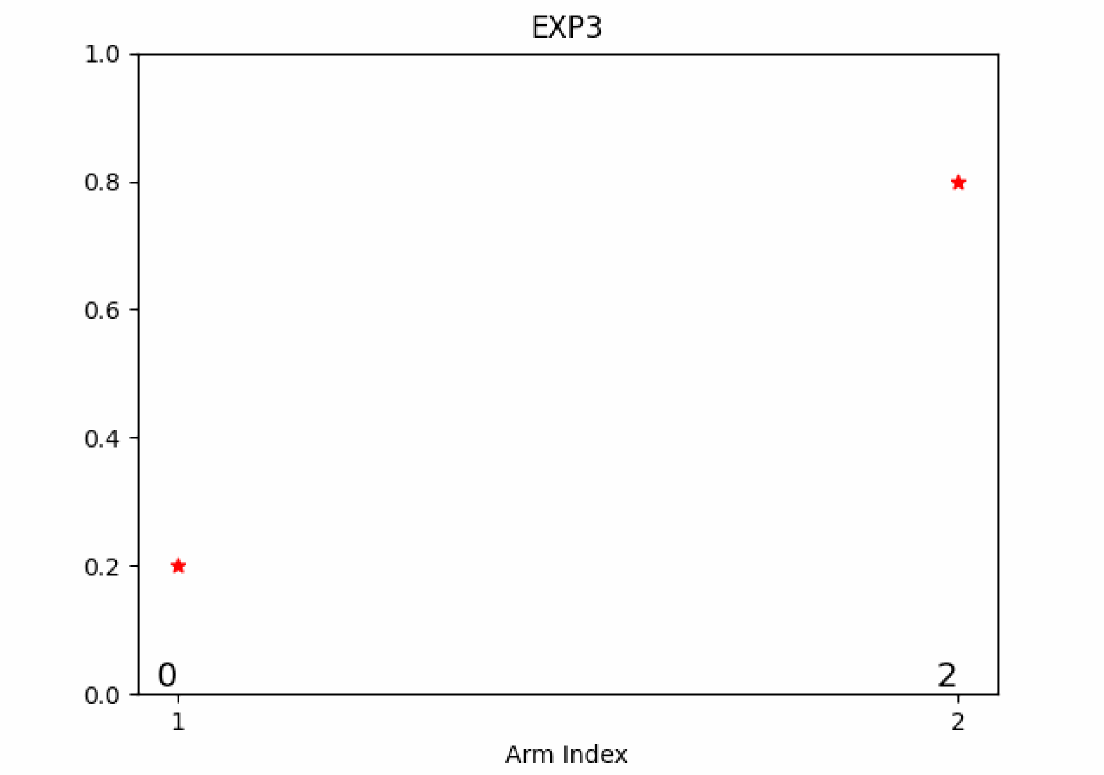
$$\hat{r}_t(a) = \frac{r_t(a)}{p_t(a)} \mathbb{I}\{a_t = a\}$$

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?

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

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

Solution 1: Adding Extra Exploration

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

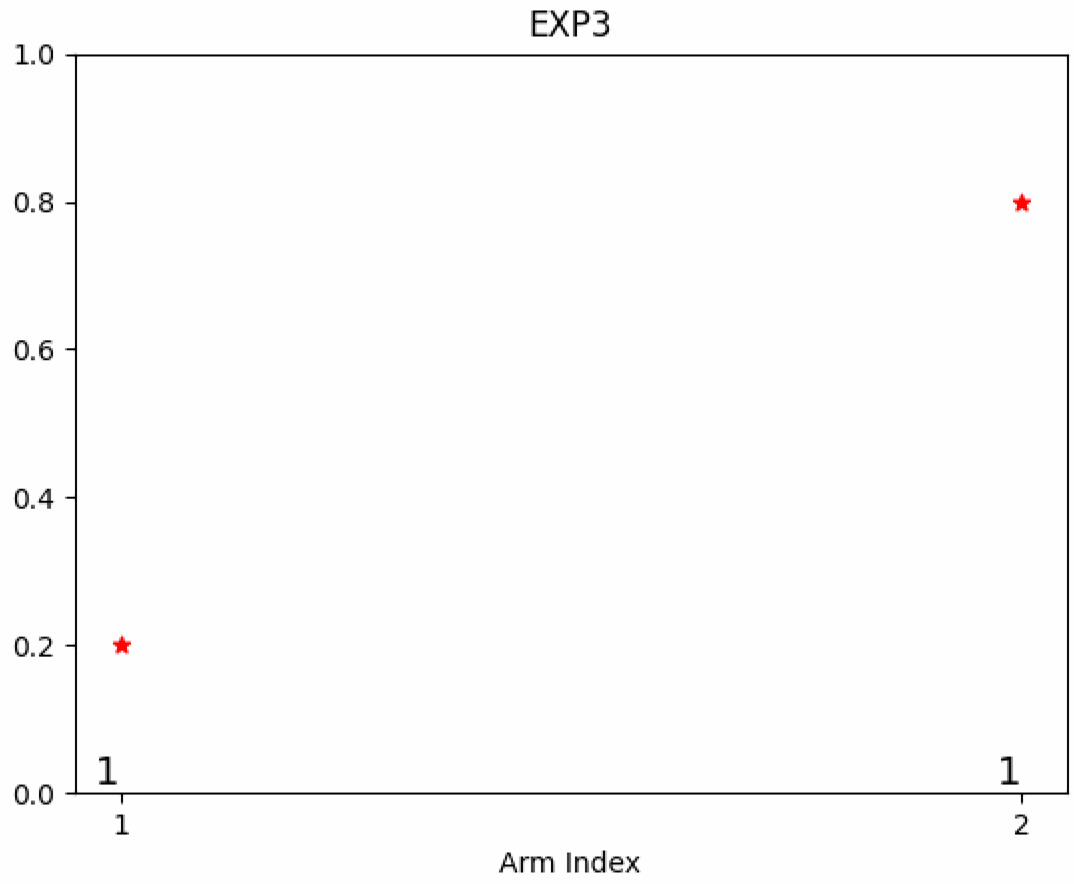
- Instead of sampling a_t according to p_t, use

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

Then the unbiased reward estimator becomes

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

Solution 1: Adding Extra Exploration



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