IE 498: Online Learning and Decision Making, Fall 2019 Sep 26th

Lecture 10: Adversarial Multi-Armed Bandits and Contextual Bandits

Lecturer: Yuan Zhou *Scribe: Rui Huang, Shulu Chen, Yuanyi Zhong*

1 Adversarial Bandit with Full Information

In the full information setting, players are able to observe all arms at time t. Given historical reward vector of arms, the player take actions and update the reward vector.In last class's example, the weight of experts' opinions affect actions. Experts with higher successful rate have higher weights. The way to update weights is that, in each run, the weight of an expert multiplies 1/2 if the action doesn't win, otherwise multiplies 1. Similarly, the probability distribution of actions in MAB Problem can be defined by weighted samples. The weights of samples can be updated after each run:

Multi-Armed Bandit (MAB) problem under Full Information Setting

at time t:

1. select arm $a_t = a$, w.p. $P_a^{(t)} := \frac{w_a^{(t)}}{w^{(t)}}$;
2. after player sees rewards, update weights $w_a^{(t)}$:

$$w_a^{(t+1)} = w_a^{(t)} \cdot \frac{2^{r_t}}{2}, \forall a$$

When action wins, $r_t = 1$, $\frac{2^{r_t}}{2} = 1$; When action loses, $r_t = 0$, $\frac{2^{r_t}}{2} = \frac{1}{2}$;
For simplification,$\frac{2^{r_t}}{2}$ can be written as 2^{r_t}, and can be further expressed as $\exp(\lambda \cdot r_{t,a})$;

$$w_a^{(t+1)} = w_a^{(t)} \cdot \exp(\lambda \cdot r_{t,a}), \forall a$$

$\lambda > 0$, when λ is large, the change of weights could be more aggressive.

Analysis:

(1) $w^{(1)} = n$; $w_a^{(1)} = 1$

(2) $\forall a, w^{(T+1)} \geq w_a^{(T+1)} = \prod_{t=1}^{T} \exp(\lambda r_{t,a}) = \exp(\lambda \sum_t r_{t,a})$.

(3) For each t:

$$\begin{aligned} w^{(t+1)} &= \sum_a w_a^{(t+1)} \\ &= \sum_a w_a^{(t)} \exp(\lambda r_{t,a}) \\ &= w^{(t)} \sum_a p_a^{(t)} \exp(\lambda r_{t,a}) \end{aligned}$$

Let $\lambda \in (0, 1]$

Since for $\forall|x| \leq 1$: $e^x \leq 1 + x + x^2$

$$\begin{aligned} w^{(t+1)} &\leq w^{(t)} \sum_a p_a^{(t)} \left(1 + \lambda r_{t,a} + \lambda^2 r_{t,a}^2\right) \\ &\leq w^{(t)} \sum_a p_a^{(t)} \left(1 + \lambda r_{t,a} + \lambda^2\right) \\ &= \left(1 + \lambda^2\right) w^{(t)} + w^{(t)} \lambda \sum_a p_a^{(t)} r_{t,a} \\ &= w^{(t)} \left(1 + \lambda^2 + \lambda \sum_a p_a^{(t)} r_{t,a}\right) \end{aligned}$$

Since $1 + x \leq e^x$

$$w^{(t+1)} \leq w^{(t)} \exp\left(\lambda^2 + \lambda \sum_a p_a^{(t)} r_{t,a}\right)$$

(4) Repeat (3) recursively:

$$\begin{aligned} w^{(T+1)} &\leq w^{(1)} \exp\left(\lambda^2 T + \lambda \sum_{t=1}^{T} \sum_a p_a^{(t)} r_{t,a}\right) \\ &= n \exp\left(\lambda^2 T + \lambda \sum_{t=1}^{T} \sum_a p_a^{(t)} r_{t,a}\right) \end{aligned}$$

$\lambda \sum_a p_a^{(t)} r_{t,a}$ is actually the expected rewards of policy.

Therefore, take log on both side:

$$\lambda \sum_a p_a^{(t)} r_{t,a} \leq \ln n + \lambda^2 T + \lambda \sum_{t=1}^{T} \sum_a p_a^{(t)} r_{t,a}$$

$$Regret = \sum_{t=1}^{T} r_{t,a} - \sum_{t=1}^{T} \sum_a p_a^{(t)} r_{t,a} \leq \frac{\ln n}{\lambda} + \lambda T \leq 2\sqrt{T \ln n}$$

Where we pick $\lambda = \sqrt{\frac{\ln n}{T}}$.

2 Bandit Feedback

In the full information setting, players could observe all arms at time t. However, in the feedback information setting, players could only observe the arm which he select. So the reward will be different. There are two ways to construct the reward $\hat{r_{t,a}}$:

$$\hat{r_{t,a}} = \frac{\mathbb{1}[a_t = a] r_t}{P_a^{(t)}} (*)$$

$$\hat{r_{t,a}} = 1 - \frac{\mathbb{1}[a_t = a](1 - r_t)}{P_a^t} (**)$$

Because $r_t \in (0,1)$, $(*) \in (0,+\infty)$, $(**) \in (-\infty,1)$. In order to use inequality $e^x \leq 1+x+x^2, (x<1)$, we choose $(**)$ as $\hat{r_{t,a}}$.
Analysis: For each time t:

$$\begin{aligned} W^{(t+1)} = \sum_a W_a^{(t+1)} &\leq W^{(t)} \sum_a P_a^{(t)}(1+\lambda \hat{r_{t,a}} + \lambda^2 \hat{r_{t,a}}^2) \\ &= W^{(t)}(1+\lambda \sum_a P_a^{(t)} \hat{r_{t,a}} + \lambda^2 \sum_a P_a^{(t)} \hat{r_{t,a}}^2) \\ &\leq W^{(t)} exp(\lambda \sum_a P_a^{(t)} \hat{r_{t,a}} + \lambda^2 \sum_a P_a^{(t)} \hat{r_{t,a}}^2) \end{aligned}$$

$$\forall a : exp(\lambda \sum_t \hat{r_{t,a}} \leq n \cdot exp(\lambda \sum_t \sum_a P_a^{(t)} \hat{r_{t,a}} + \lambda^2 \sum_t \sum_a P_a^{(t)} \hat{r_{t,a}}^2)$$

Take log:

$$\lambda \sum_t \hat{r_{t,a}} \leq \ln n + \lambda \sum_t \sum_a P_a^{(t)} \hat{r_{t,a}} + \lambda^2 \sum_t \sum_a P_a^{(t)} \hat{r_{t,a}}^2$$

Taking expection: $\mathbb{E}(\hat{r_{t,a}}) = r_{t,a}$

$$\begin{aligned} \sum_t r_{t,a} - \sum_t \sum_a P_a^{(t)} r_{t,a} &\leq \frac{\ln n}{\lambda} + \lambda \sum_t \sum_a P_a^{(t)} \mathbb{E}(\hat{r_{t,a}}^2) \\ \sum_a P_a^{(t)} \mathbb{E}(\hat{r_{t,a}}^2) &= \sum_a P_a^{(t)} \sum_{a'} P_{a'}^{(t)} [1 - \frac{\mathbb{1}(a'=a)(1-r_{a,t})}{P_a^{(t)}}]^2 \\ &\leq 1 + \sum_a (P_a^{(t)} - (1-r_{a,t}))^2 \\ &\leq n+1 \end{aligned}$$

So,

$$\begin{aligned} Regret = \sum_t r_{a,t} - \sum_t \sum_a P_a^{(t)} r_{t,a} &\leq \frac{\ln n}{\lambda} + \lambda(n+1)T \\ &\leq 2\sqrt{T(n+1)\ln n} \end{aligned}$$

3 Contextual Bandits

Now we study a new class of problems called **Contextual Bandits**. An illustration is shown below. In a contextual multi-armed bandit problem, the player has to choose between n arms. However, different from the MAB problems we have studied before, some contextual information x_t is revealed to the player before pulling arm a_t and getting reward r_t for it at time t. For example, x_t could be the user profile in recommendation systems. Different from reinforcement learning problems, the player's action doesn't affect x_t in contextual bandits.

Yet another variant of contextual bandit is the **Adversarial Contextual Bandit** which processes as follows. The adversarial setting is quite general, since the adversary can do anything about the contexts and rewards - we are not assuming any structure in the context or reward. Define a finite policy class $\Pi = \{\pi : X \mapsto A\}$, where X is the context space and A is the action space (i.e. n choices of arms $[n]$).

Adversarial Contextual Bandit
at time t:

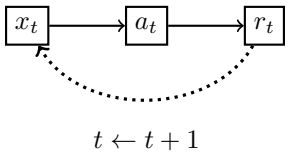

Figure 1: Illustration of contextual bandits.

1. adversary pick $x_t \in X$, $\vec{r}_t = (r_{t,1}, \ldots, r_{t,n}) \in [0,1]^{|A|}$;
2. player sees x_t, then decides a_t;
3. player receives $r_{t,a_y} = r_t$;

The **Regret** compares the cumulative reward to the best reward (in the hindsight) we can get in the policy class, which is defined as

$$R_T = \mathbb{E}\left[\max_{\pi\in\Pi}\sum_{t=1}^{T} r_{t,\pi(x_t)} - \sum_{t=1}^{T} r_t\right].$$

We again use the *multiplicative weight update* idea to develop an algorithm to solve the adversarial contextual bandit problem. This algorithm is named **EXP4** (*Exponential weighting for Exploration and Exploitation with Experts*) and was introduced in [1].

EXP4 Algorithm: Let $w^{(t)} : \Pi \mapsto \mathbb{R}_+$ be the weights of each expert (candidate policy $\pi \in \Pi$) at time t;
Set $w^{(1)}(\pi) = 1, \forall \pi \in \Pi$. For multiple rounds $t = 1, 2, \ldots$:
at time t: select $a \in A$ w.p.

$$p_a^t = \sum_{\pi:\pi(x_t)=a} \frac{w^{(t)}(\pi)}{w^{(t)}};$$

update rule (with $\lambda \in (0,1]$):

$$w^{(t+1)}(\pi) = w^{(t)}(\pi)\exp\left(\lambda\widehat{r_{t,\pi(x_t)}}\right), \forall \pi \in \Pi.$$

Remark 1. *The same unbiased reward estimator as in the previous section is used:* $\widehat{r_{t,a}} = 1 - \mathbb{1}[a_t = a]\frac{1-r_t}{p_a^{(t)}}$.

Analysis: Observations: (1) $w^{(1)} = |\Pi|$; (2) $\forall \pi, w^{(T+1)} \geq w^{(T+1)}(\pi) = \exp\left(\lambda\sum_t \widehat{r_{t,\pi(x_t)}}\right)$.

$$\begin{aligned}
w^{(t+1)} &= \sum_{\pi} w^{(t+1)}(\pi) \\
&= \sum_{a}\sum_{\pi:\pi(x_t)=a} w^{(t)}(\pi)\exp\left(\lambda\widehat{r_{t,\pi(x_t)}}\right) \\
&= w^{(t)}\sum_{a} p_a^{(t)}\exp\left(\lambda\widehat{r_{t,a}}\right) \\
&\leq w^{(t)}\sum_{a} p_a^{(t)}\left(1 + \lambda\widehat{r_{t,a}} + \lambda^2\widehat{r_{t,a}}^2\right) && \text{(since } e^x \leq 1 + x + x^2 \text{ if } x \leq 1\text{)} \\
&\leq w^{(t)}\exp\left(\lambda\sum_{a} p_a^{(t)}\widehat{r_{t,a}} + \lambda^2\sum_{a} p_a^{(t)}\widehat{r_{t,a}}^2\right) && \text{(since } 1 + x \leq e^x\text{)}
\end{aligned}$$

Therefore,

$$w^{(T+1)} \leq w^{(1)}\prod_{t=1}^{T}\exp\left(\lambda\sum_{a} p_a^{(t)}\widehat{r_{t,a}} + \lambda^2\sum_{a} p_a^{(t)}\widehat{r_{t,a}}^2\right) = |\Pi|\exp\left(\lambda\sum_{t}\sum_{a} p_a^{(t)}\widehat{r_{t,a}} + \lambda^2\sum_{t}\sum_{a} p_a^{(t)}\widehat{r_{t,a}}^2\right).$$

$$\Rightarrow \forall \pi, \sum_t \widehat{r_{t,\pi(x_t)}} - \sum_t \sum_a p_a^{(t)} \widehat{r_{t,a}} \leq \frac{\ln |\Pi|}{\lambda} + \lambda \sum_t \sum_a p_a^{(t)} \widehat{r_{t,a}}^2.$$

Take expectation (w.r.t. randomness in reward and algorithm),

$$\begin{aligned}\forall \pi, \sum_t r_{t,\pi(x_t)} - \sum_t \sum_a p_a^{(t)} r_{t,a} &\leq \frac{\ln |\Pi|}{\lambda} + \lambda \sum_t \sum_a p_a^{(t)} \mathbb{E}[\widehat{r_{t,a}}^2] \\ &\leq \frac{\ln |\Pi|}{\lambda} + \lambda T(n+1) \leq 2\sqrt{T(n+1)\ln |\Pi|}.\end{aligned}$$

Hence the **Regret Bound**: $R_T \leq \mathcal{O}\Big(\sqrt{T(n+1)\ln |\Pi|}\Big)$.

Remark 2. *There is only a* log *factor on* $|\Pi|$*, so this finite policy class can be really large, even sub-exponential in* n*, for the regret bound to be polynomial. But if* $|\Pi| > n$*, this bound is larger, because we have to compare between more policies.*

Remark 3 (Relation to Adversarial Bandit (without context))**.** *If we define the policy class to be* $n = |A|$ *constant functions* $\Pi = \{1, 2, ..., n\}$*, the problem reduces to the standard adversarial bandit. The algorithm will reduce to EXP3 and the bound is the same as before.*

References

[1] Auer, Peter, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. "The nonstochastic multiarmed bandit problem." *SIAM journal on computing* 32, no. 1 (2002): 48-77.