IE 498: Online Learning and Decision Making, Fall 2019 2019-10-01

Lecture-11: Stochastic Contextual Bandit

Lecturer: Yuan Zhou *Scribe: Junchi Yang, Siqi Zhang*

1 Settings

We set the horizon to be T, the state space X the arm space $[n] = \{1, 2, \cdots, n\}$ and policy class

$$\Pi = \{\pi : X \longrightarrow [n]\} \tag{1}$$

At round t, the player observes the state x_t, then takes action $a_t = \pi(x_t)$ and receives reward r_t, here the difference is that the reward is **stochastic** and follows a unknown distribution, i.e.

$$\left(x_t, r_{t,1}, r_{t,2}, \cdots, r_{t,n}\right) \sim \mathcal{D} \tag{2}$$

and $r_{t,k}$ denotes the reward incurred by taking arm k in round t.

We define the regret as following

$$R_T = T \cdot \max_{\pi \sim \Pi} \mathop{\mathbb{E}}_{(x, r_1, \cdots, r_n) \sim \mathcal{D}} r_{\pi(x)} - \mathbb{E} \sum_{t=1}^{T} r_t \tag{3}$$

Why are we interested in stochastic contextual bandit problem? The first point is that Exp algorithm introduced in previous lectures is expensive; the second aspect is that we consider to optimize the parameter-dependent bound of stochastic multi-arm bandit.

2 Exploration

So now the problem is that **how to learn the policy**? Our plan is that for all $\pi \in \Pi$, we estimate the following quantity

$$\mu(\pi) \triangleq \mathop{\mathbb{E}}_{(x, r_1, \cdots, r_n) \sim \mathcal{D}} r_{\pi(x)} \tag{4}$$

then we perform UCB or elimination.

Suppose we perfomr exploration with distribution $\mathcal{P}$ over Π, i.e. pick $\pi \sim \mathcal{P}$ and commit to $\pi(x_t)$. So we define

$$w_{\mathcal{P}}(x, a) \triangleq \Pr[a_t = a | x_t = x] = \sum_{\pi \in \Pi; \pi(x) = a} \mathcal{P}(\pi) \tag{5}$$

and we consider to construct the following estimator

$$\hat{\mu}(\pi) \sim \mathop{\mathbb{E}}_{x \sim D} \hat{r}_{\pi(x)} \tag{6}$$

first we construct an unbiased estimator for each arm

$$\hat{r}_{t,a} = \frac{r_t \mathbb{1}[a_t = a]}{w_P(x,a)} \tag{7}$$

and we define

$$\hat{\mu}_t(\pi) \triangleq \hat{r}_{t,\pi(x_t)}, \quad \hat{\mu}_{\mathcal{Z}}(\pi) \triangleq \frac{1}{|\mathcal{Z}|} \sum_{t \in \mathcal{Z}} \hat{r}_{t,\pi(x_t)} \tag{8}$$

so we have that

$$\hat{\mu}(\pi) = r_{t,\pi(x_t)} \Big(\frac{1}{|\mathcal{Z}|} \sum_{t \in \mathcal{Z}} \hat{r}_{t,\pi(x_t)} \Big) \tag{9}$$

3 Regret Analysis

The proposed method is **modified exploration**:

Algorithm 1 Modified Exploration

for $\lambda \in (0, \frac{1}{2}]$, do exploration following

$$\begin{cases} \text{pick } \pi \in P, \text{play } a = \pi(x_t) & \text{w.p. } 1-\lambda \\ \text{pick } a \sim \text{Unif}([n]) & \text{w.p. } \lambda \end{cases} \tag{10}$$

Then set the probability of choosing arm a in modified exploration to be

$$w'_P(x,a) \triangleq (1-\lambda) w_P(x,a) + \frac{\lambda}{n} = \Pr[a_t = a | x_t = x] \tag{11}$$

Lemma 1 (**Bernstein Inequality**). *Let $X_1, ..., X_n$ be independent real-valued random variables. If $X_i \in [0, M] \quad \forall i$, then*

$$\mathbb{P}\left[\left| \sum_{i=1}^n X_i - \mathbb{E} \sum_{i=1}^n X_i \right| > t \right] \leq 2 \exp\left(- \frac{t^2/2}{\sum_{i=1}^n \text{Var}(X_i) + Mt/3} \right) \tag{12}$$

So we have

$$\text{Var}[\hat{r}_{t,a}] \leq \mathbb{E}_{a,t} \hat{r}^2_{t,a} = \sum_{a'} w_P(x_t, a') \cdot \frac{r_t^2 \mathbb{1}[a'_t = a]}{w_P(x_t, a)} = \frac{r_t^2}{w_P(x_t, a)} \tag{13}$$

Lemma 2 (**Minmax Theorem**). *Let X, Y be compact sets and $f : X \times Y \longrightarrow \mathbb{R}$ be continuous. If $f(\cdot, y)$ is convex $\forall y$ and $f(x, \cdot)$ is concave $\forall x$, then*

$$\min_{x \in X} \max_{y \in Y} f(x,y) = \max_{y \in Y} \min_{x \in X} f(x,y) \tag{14}$$

Theorem 3. *There exists a distribution P over Π such that using modified exploration associated with P as defined in (10) and $\lambda > \epsilon/3$, we have*

$$\mathbb{P}\left[|\hat{\mu}_Z(\pi) - \mu(\pi)| > \epsilon \right] \leq 2 exp\left(- \frac{\epsilon^2 |Z|}{6n} \right), \quad \forall \pi \in \Pi. \tag{15}$$

Proof. Since $\hat{\mu}_Z(\pi) = \frac{1}{|Z|}\sum_{t\in Z}\hat{r}_{t,\pi(x_t)}$ and $\hat{\mu}_Z(\pi)$ is an unbiased estimator for $\mu(\pi)$, we can use Bernstein Inequality to derive this tail bound. Therefore, we need to find M such that $\hat{r}_{t,\pi(x_t)} \le M$ and an upper bound for $\mathrm{Var}[\hat{r}_{t,\pi(x_t)}]$. Note that

$$\hat{r}_{t,a} = r_t \cdot 1_{[a_t=a]}/w'_P(x_t,a) \le \frac{1}{\lambda/n} = n/\lambda \tag{16}$$

So we let $M = n/\lambda$, and now it is left to show there exists a distribution $P \in \Delta_\Pi$ (Δ_Π is the class of all distributions over Π) such that $\mathrm{Var}[\hat{r}_{t,\pi(x_t)}] \le 2n$. Note that

$$\operatorname*{Var}_{(x_t,\vec{r}_t)\sim D, a_t\sim w'_P}[\hat{r}_{t,\pi(x_t)}] \le \operatorname*{\mathbb{E}}_{D,w'_P}[(\hat{r}_{t,\pi(x_t)})^2] \le \operatorname*{\mathbb{E}}_{x_t\sim D}\left[\frac{1}{w'_P(x_t,\pi(x_t)}\right] \le \max_{Q\in\Delta_\Pi}\operatorname*{\mathbb{E}}_{\pi\sim Q, x\sim D}\frac{1}{w'_P(x,\pi(x))} \tag{17}$$

Recall that $w'_P(x,\pi(x)) = (1-\lambda)\mathbb{E}_{\theta\sim P}1_{[\pi(x)=\theta(x)]} + \frac{\lambda}{n}$, and define

$$f(P,Q) = \operatorname*{\mathbb{E}}_{x\sim D}\operatorname*{\mathbb{E}}_{\pi\sim Q}\left[\left((1-\lambda)\mathbb{E}_{\theta\sim P}1_{[\pi(x)=\theta(x)]} + \frac{\lambda}{n}\right)^{-1}\right] \tag{18}$$

It can be verified that $f(P,Q)$ is linear in Q and convex in P, so by Minmax Theorem and equation (17), we have

$$\min_{P\in\Delta_\Pi}\operatorname*{Var}_{D,w'_P}[\hat{r}_{t,\pi(x_t)}] \le \max_{Q\in\Delta_\Pi}\min_{P\in\Delta_\Pi} f(P,Q) \le \max_{Q\in\Delta_\Pi} f(Q,Q). \tag{19}$$

where

$$f(Q,Q) = \mathbb{E}_{x\sim D}\sum_a w_Q(x,a)\left((1-\lambda)w_Q(x,a) + \frac{\lambda}{n}\right)^{-1} \le \mathbb{E}_{x\sim D}\sum_a \frac{1}{1-\lambda} \le 2n$$

therefore,

$$\min_{P\in\Delta_\Pi}\operatorname*{Var}_{D,w'_P}[\hat{r}_{t,\pi(x_t)}] \le 2n. \tag{20}$$

So we conclude there exists a distribution $P \in \Delta_\Pi$ such that $\mathrm{Var}[\hat{r}_{t,\pi(x_t)}] \le 2n$. Combining with $\hat{r}_{t,a} \le n/\lambda$ and plugging into Bernstein Inequality, we have

$$\begin{aligned}
\mathbb{P}\left[|\hat{\mu}_Z(\pi) - \mu(\pi)| > \epsilon\right] \le & 2\exp\left(-\frac{\epsilon^2|z|^2/2}{2n|z| + \frac{1}{3}\frac{n}{\lambda}\epsilon|z|}\right) \\
\le & 2\exp\left(-\frac{\epsilon^2|z|^2/2}{2n + \frac{1}{3}\frac{n}{\lambda}\epsilon}\right) \\
\le & 2\exp\left(-\frac{\epsilon|z|/2}{6n}\right), \quad \text{when} \quad \lambda > \epsilon/3
\end{aligned}$$

□

So the desired size of Z is

$$|Z| \sim \frac{n}{\epsilon^2}\log(|\Pi|) \tag{21}$$