IE 498: Online Learning and Decision Making, Fall 2019 Sep 19th

Lecture 08: Two and Multi-Armed Bandit

Lecturer: Yuan Zhou *Scribe: Zexing Xu, Erik Pan*

1 Recap

$\exists A : (\epsilon, \delta) - PAC$ for the better arm, sample complexity $\leq T \Rightarrow \Delta(D_1, D_2) \geq 1 - 2\delta$, where

$$D_1 = B_p^{\otimes T} \otimes B_{p+\epsilon}^{\otimes T}$$

$$D_2 = B_{p+\epsilon}^{\otimes T} \otimes B_p^{\otimes T}$$

, where Δ is total variation distance, D_1, D_2 are two different distributions defined on the same probability space and B_p is Bernoulli distribution with parameter p.

Pinsker's inequality (proved in Lecture 07) stated

$$\textcircled{1} : \Delta(P, Q) \leq \sqrt{D_{KL}(P||Q)}$$
$$\textcircled{2} : \Delta(P, Q) \leq 1 - \frac{1}{2} exp(-D_{KL}(P||Q))$$

If D_{KL} is small (i.e. $\approx \epsilon$) then

$$\textcircled{1} : \Delta(P, Q) \leq \sqrt{\epsilon}$$
$$\textcircled{2} : \Delta(P, Q) \leq 1 - \frac{1}{2} e^{-\epsilon}$$

So ① gives a better bound.

While if D_{KL} is large (i.e. ≈ 100) then

$$\textcircled{1} : trivial$$
$$\textcircled{2} : \Delta(P, Q) \leq 1 - \frac{1}{2} e^{-100} < 1$$

So ② gives a better bound.

2 Two-Armed Bandit

As discussed in Lecture 7, K-L Divergence has some limitations, but why it is still frequently used in measuring the variational difference? The answer lies in the following promising properties of K-L Divergence.

Fact *(Additivity)*: Suppose $P(x, y) = P_1(x)P_2(y), Q(x, y) = Q_1(x)Q_2(y)$, then :

$$\begin{aligned}
D_{KL}(P,Q) &= -\sum_{x,y} P(x,y)\cdot ln(\frac{Q(x,y)}{P(x,y)}) \\
&= -\sum_{x,y} P_1(x)P_2(y)[ln(\frac{Q_1(x)}{P_1(x)}) + ln(\frac{Q_2(y)}{P_2(y)})] \\
&= -\sum_{x} P_1(x)[ln(\frac{Q_1(x)}{P_1(x)})] - \sum_{y} P_2(y)[ln(\frac{Q_2(y)}{P_2(y)})] \\
&= D_{KL}(P_1||Q_1) + D_{KL}(P_2||Q_2)
\end{aligned}$$

Since D_{KL} is not symmetric, we can't combine the 2 terms, so we have :

$$D_{KL}(D_1||D_2) = T\cdot D_{KL}(B_p||B_{p+\epsilon}) + T\cdot D_{KL}(B_{p+\epsilon}||B_p)$$

Since :

$$\begin{aligned}
D_{KL}(B_p||B_{p+\epsilon}) &= -p\cdot ln(1+\frac{\epsilon}{p}) - (1-p)\cdot ln(1+\frac{\epsilon}{1-p}) \\
&= -p(\frac{\epsilon}{p} - \frac{1}{2}\frac{\epsilon^2}{p^2} + \mathcal{O}(\frac{\epsilon^3}{p^3})) - (1-p)(-\frac{\epsilon}{1-p} - \frac{1}{2}\frac{\epsilon^2}{(1-p)^2} + \mathcal{O}\frac{\epsilon^3}{(1-p)^3}) \ \ (Taylor\ Expansion) \\
&= \epsilon^2\cdot(\frac{1}{2p} + \frac{1}{2\cdot(1-p)}) + \mathcal{O}(\epsilon^3)\cdot(\frac{1}{p^3} + \frac{1}{(1-p)^3}) \\
&= \mathcal{O}(\epsilon^3) \ \ (as\ long\ as\ p\ is\ away\ from\ 0,1\ (e.g.\ p\in[0.1,0.9])\ and\ \epsilon\ is\ small)
\end{aligned}$$

So :

$$D_{KL}(D_1||D_2) = \mathcal{O}(\epsilon^2\cdot T)$$

So by Pinster's inequality ① :

$$\Delta(D_1,D_2) \le \mathcal{O}(\epsilon\cdot\sqrt{T})$$

when $\delta > 0.01$,

$$\left.\begin{aligned}\Delta(D_1,D_2) &\ge 0.9 \\ \Delta(D_1,D_2) &\le \mathcal{O}(\epsilon\cdot\sqrt{T})\end{aligned}\right\} \Rightarrow \epsilon\cdot\sqrt{T} \ge \Omega(1) \Rightarrow T \gtrsim \frac{1}{\epsilon^2}$$

For constant δ, the algorithm give us $T \lesssim \frac{1}{\epsilon^2} log(\frac{1}{\delta})$, and δ is constant, the two bounds matches.

when $\delta << 0.01$ (i.e. δ is very small), by Pinster's inequality ② :

$$\left.\begin{aligned}\Delta(D_1,D_2) &\le 1 - \frac{1}{2}e^{-\Theta(T\epsilon^2)} \\ \Delta(D_1,D_2) &\ge 1-2\delta\end{aligned}\right\} \Rightarrow e^{-\Theta(T\epsilon^2)} \le 4\delta \Rightarrow -\Theta(T\epsilon^2) \le ln(4\delta) \Rightarrow T\epsilon^2 \gtrsim ln(\frac{1}{\delta}) \Rightarrow T \gtrsim \epsilon^{-2} ln(\frac{1}{\delta})$$

3 Multi-Armed Bandit

For any bandit instance, let $I = (D_1, D_2, ..., D_n)$ be the reward distribution for each arm i, and (possibly randomize) policy π, running π on I result the history $H = (a_1, r_1, a_2, r_2, ..., a_{|H|}, r_{|H|})$ where $|H|$ is the length of H. Let $P_{\pi,I}$ be the distribution of H.

Lemma 1. *Let* $I = (D_1, D_2, ..., D_n), I' = (D'_1, D'_2, .., D'_n)$ *for any* π *:*

$$D_{KL}(P_{\pi,I}||P_{\pi,I'}) = \sum_{i=1}^{n} \mathop{\mathbb{E}}_{H\sim P_{\pi,I}}[T_i(H)]\cdot D_{KL}(D_i||D'_i) \ \ (\textit{where } T_i[H] \textit{ is how many times arm } i \textit{ is played in } H)$$

Lemma 1 describes the relationship between the whole distribution and the original individual distributions. Intuitively, if we run K-L divergence on similar distance, K-L does not diverge much.

Proof.

$$\begin{aligned} P_{\pi,I}(H) &= P_\pi(a_1)\cdot P_{D_{a_1}}(r_1)\cdot P_\pi(a_2)\cdot P_{D_{a_2}}(r_2)\cdot P_\pi(a_3)\cdot P_{D_{a_3}}(r_3)\cdot ... \\ &= \prod_{t=1}^{|H|} P_\pi(a_t|a_1, r_1, a_2, r_2, ..., a_{t-1}, r_{t-1})\cdot P_{D_{a_t}}(r_t) \end{aligned}$$

Similarly, we have

$$P_{\pi,I'}(H) = \prod_{t=1}^{|H|} P_\pi(a_t|a_1, r_1, a_2, r_2, ..., a_{t-1}, r_{t-1})\cdot P_{D'_{a_t}}(r_t)$$

For $P_{\pi,I'}(H) \neq 0$, we have

$$ln(\frac{P_{\pi,I}(H)}{P_{\pi,I'}(H)}) = \sum_{t-1}^{|H|} ln(\frac{P_{D_{a_t}}(r_t)}{P'_{D_{a_t}}(r_t)})$$

Then it follows that

$$D_{KL}(P_{\pi,I}(H)||P_{\pi,I'}(H)) = \mathop{\mathbb{E}}_{H\sim P_{\pi,I}} ln(\frac{P_{\pi,I}(H)}{P_{\pi,I'}(H)}) = \mathop{\mathbb{E}}_{H\sim P_{\pi,I}} \sum_{t=1}^{|H|} ln(\frac{P_{D_{a_t}}(r_t)}{P'_{D_{a_t}}(r_t)})$$

$$\begin{aligned} \sum_{t=1}^{\infty} \mathop{\mathbb{E}}_{H\sim P_{\pi,I}} \mathbb{I}[t\leq|H|] ln(\frac{P_{D_{a_t}}(r_t)}{P'_{D_{a_t}}(r_t)}) &= \sum_{t=1}^{\infty} \mathop{\mathbb{E}}_{H\sim P_{\pi,I}} \mathbb{I}[t\leq|H|] \sum_{i=1}^{n} \mathbb{I}[a_t = i] ln(\frac{D_i(r_t)}{D'_i(r_t)}) \ \ (\textit{where } \mathbb{I}[.] \textit{ is indicator function}) \\ &= \sum_{t=1}^{\infty}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{(a_1,t_1,...,a_{t-1},r_{t-1})\sim P_{\pi,I}} \mathbb{I}[a_t \textit{ exist} \wedge a_t = i] \mathop{\mathbb{E}}_{r_t\sim D_i} ln(\frac{D_i(r_t)}{D'_i(r_t)}). \\ &= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{H\sim P_{\pi,I}} [T_i(H)] - D_{KL}(D_i||D'_i) \end{aligned}$$

Note: Since a_t and r_t is independent of H, we do not need conditional expectation for the term $\mathrm{E}_{r_t\sim D_i} ln(\frac{D_i(r_t)}{D'_i(r_t)})$. □

Theorem 2 (minimax regret lower bound)**.** $\forall n \geq 2, T \geq n$*, policy* π*,* $\exists\ I$ *such that* $R_T^{\pi,I} \gtrsim \sqrt{nT}$

Proof. Let $I = (B_{\frac{1}{2}+\Delta}, B_{\frac{1}{2}}, B_{\frac{1}{2}}, ..., B_{\frac{1}{2}})$, let $Z = \mathop{\mathrm{argmin}}_{i\in\{2,3,...,n\}} \mathop{\mathbb{E}}_{H\sim P_{\pi,I}} [T_i(H)]$.
Set $I' = (B_{\frac{1}{2}+\Delta}, B_{\frac{1}{2}}, B_{\frac{1}{2}}, ..., B_{\frac{1}{2}}, B_{\frac{1}{2}+2\Delta}, B_{\frac{1}{2}}, B_{\frac{1}{2}}, ..., B_{\frac{1}{2}})$. Typically, we set the reward distribution for z th arm as $B_{\frac{1}{2}+2\Delta}$ to have slightly better performance than our original best arm1, but this arm may not receive

much attention.
Let A be the event $\{T_1(H) \leq \frac{T}{2}\}$ (Play 1st arm less than 1/2 of the time), clearly:

$$\begin{cases} R_T^{\pi,I}|_A \geq \frac{T}{2}\Delta \\ R_T^{\pi,I'}|_{\bar{A}} \geq \frac{T}{2}\Delta \end{cases}$$

$$\begin{aligned} R_T^{\pi,I} + R_T^{\pi,I'} &\geq \frac{T\Delta}{2} \cdot \Pr_{\pi,I}[A] + \Pr_{\pi,I'}[\bar{A}]) = (\Pr_{\pi,I}[A] + (1 - \Pr_{\pi,I'}[A])) = \frac{T\Delta}{2} \cdot (1 - |\Pr_{\pi,I}[A] - \Pr_{\pi,I'}[A]|) \\ &\geq \frac{T\Delta}{2}(1 - \delta(P_{\pi,I}, P_{\pi,I'})) \\ &\geq \frac{T\Delta}{2}(1 - \sqrt{D_{KL}(P_{\pi,I}||P_{\pi,I'})}\ ) \ (By\ Pinsker's\ Inequality\ \textcircled{1}) \end{aligned}$$

Since

$$D_{KL}(P_{\pi,I}||P_{\pi,I'}) = \mathop{\mathbb{E}}_{\pi,I}[T_z] D_{KL}(D_Z, D'_Z) \leq \frac{T}{n-1}(2\Delta^2) \lesssim \frac{T}{n-1}\Delta^2$$

So

$$D_{KL}(P_{\pi,I}||P_{\pi,I'}) \geq \frac{T\delta}{2}(1 - \mathcal{O}(\Delta\sqrt{\frac{T}{n}}))$$

Pick $\delta = c \cdot \sqrt{\frac{n}{T}}$, where c is any constant such that $(1 - \mathcal{O}(\Delta\sqrt{\frac{T}{n}})) < \frac{1}{2}$

$$D_{KL}(P_{\pi,I}||P_{\pi,I'}) \geq \frac{T \cdot c\sqrt{\frac{n}{T}}}{4} \gtrsim \sqrt{nT}$$

□

Similarly, we can proof PAC has a lower bound of $n\epsilon^{-2}ln(\frac{1}{\delta})$