IE613: Online Machine Learning **Jan-Apr 2018**

Lecture 24: Lower Bound for Stochastic MAB

Lecturer: M. K. Hanawal *Scribes: Nimita Shinde*

24.1 Recap

In the last lecture, we saw different algorithms with Bayesian and frequentist approach and the upper bound on the regret for each of the algorithms. All the algorithms have an upper bound on regret that is $\mathcal{O}(logT)$, where T is the time horizon.

In this lecture, we look at the lower bound on the regret of stochastic Multi-Armed Bandit (MAB) problem. We will show that the bound obtained in the previous lectures is unimprovable i.e. the regret will at least be of the order $log(T)$.

Note: The expression $log(\cdot)$ used in these notes refers to natural logarithm i.e. $log_e(\cdot)$.

24.2 Lower Bound for Stochastic MAB

In the adversarial setting, in order to find a lower bound, we created a tree and generated a sequence on which algorithm makes at least *some* number of mistakes. This number was then used to lower bound the number of mistakes. In the stochastic MAB setting, we find the minimum number of pulls of sub-optimal arms that an algorithm has to make for any given sequence in the process of finding the optimal arm.

In this section, we focus on those policies/strategies that eventually learn (they can find an optimal arm as $T \to \infty$). These policies are called admissible policies. Their definition is given below.

Definition 24.1 (Admissible Policy/Strategy) *A strategy is admissible if* $\forall i \neq i^*, \Delta_i > 0$*, we have,*

$$\mathbb{E}[N_i(T)] = o(T^\alpha) \ \forall \alpha > 0.$$

Here, $i \in [K]$ is the set of all arms in the MAB setting, i^* is the optimal arm, and $N_i(T)$ defines the number of times arm i is played till time T. μ_i is the expected value of arm i, $\mu^* \geq \max_{i\in[K]} \mu_i$ refers to the expected value of the optimal arm and $\Delta_i = \mu^* - \mu_i$.

Next, we give a theorem that states a lower bound on the number of plays of suboptimal arm in a Bernoulli game and derive the theorem.

Theorem 24.2 *Consider a stochastic MAB with any set of Bernoulli reward distributions. Then for any suboptimal arm* i $(\Delta_i > 0)$*, the following holds*

$$\liminf_{T\to\infty} \frac{N_i(T)}{log(T)} \geq \frac{1}{D(\mu_i, \mu^*)} \tag{24.1}$$

where $D(\mu_i, \mu^*) = \mu_i log\frac{\mu_i}{\mu^*} + (1-\mu_i)log\frac{1-\mu_i}{1-\mu^*}$ *is the KL divergence between a Bernoulli of parameter* μ_i *and a Bernoulli of parameter* μ^*.

Proof: For simplicity, we only consider the case of two arms in this proof. It is organized in three steps.
Step I: Notations

We have assumed $K = 2$. Without loss of generality, assume that arm 1 is the optimal arm and arm 2 is suboptimal in the bandit problem under consideration. Consider two bandit problem instances as:

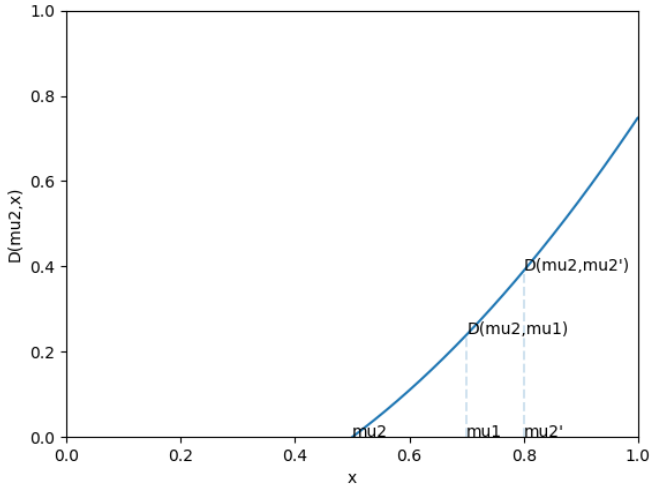
$$Bandit\ Instance\ 1 : (\mu_1, \mu_2),\ \mu_1 > \mu_2$$
$$Bandit\ Instance\ 2 : (\mu_1, \mu_2'),\ \mu_1 < \mu_2'$$

The instance 1 is the bandit problem under consideration. But we can setup another problem instance with μ_2' selected in such a way that the player cannot distinguish between instance 1 and 2. In instance 2, arm 1 is the suboptimal arm. Thus, incorrect detection of optimal arm would lead to loss of reward. In this situation, the player will have to pull both the arms certain number of times to detect the correct problem instance. Formally, we can define this as: given $\epsilon > 0$, $1 > \mu_1 > \mu_2$, there exists $\mu_2' \in (\mu_1, 1)$ such that

$$D(\mu_2, \mu_2') \leq (1+\epsilon)D(\mu_2, \mu_1). \tag{24.2}$$

We know that $D(\mu_2, x)$ for a given μ_2 is strictly convex in x and increasing in the interval $x \in [\mu_2, 1]$. Also, $\mu_2 < \mu_1 < \mu_2'$ by virtue of their definition. Thus, $D(\mu_2, \mu_2')$ is always greater than $D(\mu_2, \mu_1)$ and we can find an ϵ that satisfies the above equation. This is also evident from the graph given below.

Figure 24.1: 'Divergence'



Let us define some of the notations that are used in the proof:

- Let $\mathbb{E}'$ and $\mathbb{P}'$ be the expectation and probability respectively with respect to the second bandit instance. Using this, we compare the behavior of the player on the initial instance (1) and modified bandit instance (2). In particular, we prove that the player cannot distinguish between the two problem instances with non-negligible probability. In this case, since we have an admissible policy, the algorithm

will not make large number of mistakes on the modified instance. So, we have a lower bound on the number of times an optimal arm in played in instance 2. This reasoning implies a lower bound on the number of times arm 2 is played in instance 1.

- For $s \in \{1, ..., T\}$, let
$$\hat{D}_s = \sum_{t=1}^{s} log \frac{\mu_2 X_{2,t} + (1-\mu_2)(1-X_{2,t})}{\mu_2' X_{2,t} + (1-\mu_2')(1-X_{2,t})} \tag{24.3}$$
where $\{X_{2,t}\}_{t=1,...s}$ is the sequence of samples observed from arm 2 ($X_{2,i}$ is the i^{th} observed sample of arm 2. $\{X_{2,t}\}$ forms i.i.d. sequence from a Bernoulli of parameter μ_2.
$$\begin{aligned}\mathbb{E}_X\left[log \frac{\mu_2 X_{2,t} + (1-\mu_2)(1-X_{2,t})}{\mu_2' X_{2,t} + (1-\mu_2')(1-X_{2,t})}\right] &= Pr\{X_{2,t}=1\} log \frac{\mu_2}{\mu_2'} + Pr\{X_{2,t}=0\} log \frac{1-\mu_2}{1-\mu_2'} \\ &= \mu_2 log \frac{\mu_2}{\mu_2'} + (1-\mu_2) log \frac{1-\mu_2}{1-\mu_2'} \\ &= D(\mu_2, \mu_2')\end{aligned}$$
So, we can derive $\mathbb{E}[\hat{D}_s]$ as,
$$\begin{aligned}\mathbb{E}[\hat{D}_s] &= \mathbb{E}\left[\sum_{t=1}^{s} log \frac{\mu_2 X_{2,t} + (1-\mu_2)(1-X_{2,t})}{\mu_2' X_{2,t} + (1-\mu_2')(1-X_{2,t})}\right] \\ &= \sum_{t=1}^{s} \mathbb{E}\left[log \frac{\mu_2 X_{2,t} + (1-\mu_2)(1-X_{2,t})}{\mu_2' X_{2,t} + (1-\mu_2')(1-X_{2,t})}\right] \\ &= \sum_{t=1}^{s} D(\mu_2, \mu_2') \\ &= sD(\mu_2, \mu_2')\end{aligned}$$
Thus, we can say that $\hat{D}_{N_2(T)}$ is the non-renormalized empirical estimate of $D(\mu_2, \mu_2')$ at time T.
- For any event $\mathcal{A}$ in the σ-algebra generated by $X_{2,1}, ..., X_{2,N_2(T)}$, the following change-of-measure identity holds:
$$\mathbb{P}'(\mathcal{A}) = \mathbb{E}[\mathbf{1}_{\mathcal{A}} exp(-\hat{D}_{N_2(T)})] \tag{24.4}$$
- Define an event as:
$$C_T = \left\{N_2(T) < \frac{1-\epsilon}{D(\mu_2, \mu_2')} log(T) \text{ and } \hat{D}_{N_2(T)} \leq \left(1 - \frac{\epsilon}{2}\right) log(T)\right\} \tag{24.5}$$

Step II: $\mathbb{P}(C_T) = o(1)$

In this step, we show the above inequality.

From equations 24.4 and 24.5, we can write,
$$\begin{aligned}\mathbb{P}'(C_T) &= \mathbb{E}[\mathbf{1}_{C_T} exp(-\hat{D}_{N_2(T)})] \\ &\geq \mathbb{E}[\mathbf{1}_{C_T} exp(-(1-\epsilon/2) log(T)) \\ &= \mathbb{E}[\mathbf{1}_{C_T}] T^{-(1-\epsilon/2)} \\ &= \mathbb{P}(C_T) T^{-(1-\epsilon/2)}\end{aligned}$$
Let
$$f_T = \frac{1-\epsilon}{D(\mu_2, \mu_2')} log(T)$$

Continuing with the above derivation and using equation 24.5 and Markov's inequality, the following is true:

$$\begin{aligned}\mathbb{P}(C_T) &\leq \mathbb{P}'(C_T)T^{(1-\epsilon/2)} \\ &\leq \mathbb{P}'(N_2(T) < f_T)T^{(1-\epsilon/2)} \\ &= \mathbb{P}'(T - N_2(T) > T - f_T)T^{(1-\epsilon/2)} \\ \mathbb{P}(C_T) &\leq \frac{\mathbb{E}'[T - N_2(T)]}{T - f_T}T^{(1-\epsilon/2)}\end{aligned}$$

Note that, in the term $\frac{T^{(1-\epsilon/2)}}{T-f_T}$ from the above equation, the numerator has a power of T strictly less than 1, $((1-\epsilon/2) < 1)$, and denominator is linearly proportional to T. Therefore, as $T \to \infty$, $\frac{T^{1-\epsilon/2}}{T-f_T} \to 0$. Also, note that, in the bandit instance 2, arm 2 is the optimal arm. Thus, with the admissible policy assumption, we can say that for large enough T, more and more number of plays will be of the optimal arm 2 for instance 2. Thus, $\mathbb{E}'[T - N_2(T)]$ will tend to zero. This implies that,

$$\mathbb{P}(C_T) \leq \frac{\mathbb{E}'[T - N_2(T)]}{T - f_T}T^{1-\epsilon/2} = o(1). \tag{24.6}$$

Step III: $\mathbb{P}(N_2(T) < f_T) = o(1)$

In this step, we prove the above inequality.

If $N_2(T) < f_T$, then note that the event, $\Big\{\hat{D}_{N_2(T)} \leq \big(1 - \frac{\epsilon}{2}\big)log(T)\Big\}$ is a strict subset of the event $\Big\{\max_{s\leq f_T} \hat{D}_s \leq \big(1 - \frac{\epsilon}{2}\big)log(T)\Big\}$. Thus, we can say that,

$$\begin{aligned}\mathbb{P}(C_T) &\geq \mathbb{P}\Big[N_2(T) < f_T \ and \ \max_{s\leq f_T} \hat{D}_s \leq (1 - \frac{\epsilon}{2})log(T)\Big] \\ &= \mathbb{P}\Big[N_2(T) < f_T \ and \ \frac{D(\mu_2, \mu_2')}{(1-\epsilon)log(T)} \max_{s\leq f_T} \hat{D}_s \leq \frac{1-\epsilon/2}{1-\epsilon}D(\mu_2, \mu_2')\Big]\end{aligned} \tag{24.7}$$

The second equality is obtained by multiplying the second event by $\frac{D(\mu_2,\mu_2')}{(1-\epsilon)}$ on both sides. Since the divergence and the term $(1-\epsilon)$ are both non-negative, the inequality still holds.

Law of large numbers is given as,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} X_t = \mu \ a.s.$$

The maximal version of strong law of large numbers can be defined as: for any sequence $\{X_t\}$ of independent real random variables with positive mean $\mu > 0$, the law of large numbers implies that

$$\lim_{n\to\infty} \frac{1}{n} \max_{s=1,\ldots,n} \sum_{t=1}^{s} X_t = \mu \ a.s.$$

Let us consider second term of the event defined in the right hand side of equation 24.7. Since $D(\mu_2, \mu_2') > 0$ and $\frac{1-\epsilon/2}{1-\epsilon} > 1$, we can say that,

$$\mathbb{P}\Big[\frac{D(\mu_2, \mu_2')}{(1-\epsilon)log(T)} \max_{s\leq f_T} \hat{D}_s \leq \frac{1-\epsilon/2}{1-\epsilon}D(\mu_2, \mu_2')\Big] \geq \mathbb{P}\Big[\frac{D(\mu_2, \mu_2')}{(1-\epsilon)log(T)} \max_{s\leq f_T} \hat{D}_s \leq D(\mu_2, \mu_2')\Big]$$

Note that $\frac{D(\mu_2,\mu_2')}{(1-\epsilon)log(T)} = \frac{1}{f_T}$. As $T \to \infty$, the term $log(T)$ dominates, thus $\frac{1}{f_T} \to 0$. So,

$$\begin{aligned}&\lim_{T\to\infty} \mathbb{P}\Big[\frac{D(\mu_2, \mu_2')}{(1-\epsilon)log(T)} \max_{s\leq f_T} \hat{D}_s \leq D(\mu_2, \mu_2')\Big] = 1 \\ \Rightarrow &\lim_{T\to\infty} \mathbb{P}\Big[\frac{D(\mu_2, \mu_2')}{(1-\epsilon)log(T)} \max_{s\leq f_T} \hat{D}_s \leq \frac{1-\epsilon/2}{1-\epsilon}D(\mu_2, \mu_2')\Big] = 1\end{aligned}$$

So, as $T \to \infty$, the probability of second event on the right hand side of equation 24.7 is 1, it will always occur, so we can write 24.7 as,

$$\lim_{T\to\infty} \mathbb{P}(C_T) \geq \lim_{T\to\infty} \mathbb{P}\Big[N_2(T) < f_T\Big]$$
$$o(1) = \lim_{T\to\infty} \mathbb{P}\Big[N_2(T) < f_T\Big]$$

The last inequality is derived from the result of Step II. As $\lim_{T\to\infty} \mathbb{P}[N_2(T) < f_T] = o(1)$, $\lim_{T\to\infty} \mathbb{P}[N_2(T) \geq f_T] \approx 1$. So, $\lim_{T\to\infty} \mathbb{E}[N_2(T)] \geq (1 + o(1))f_T$. Replacing the value of f_T,

$$\lim_{T\to\infty} \mathbb{E}[N_2(T)] \geq (1 + o(1))\frac{(1-\epsilon)log(T)}{D(\mu_2, \mu_2')}$$
$$\geq (1 + o(1))\frac{(1-\epsilon)}{(1+\epsilon)}\frac{log(T)}{D(\mu_1, \mu_2)}$$

The second inequality follows since $D(\mu_2, \mu_2') \leq (1 + \epsilon)D(\mu_1, \mu_2)$. So,

$$\Rightarrow \lim_{T\to\infty} \frac{\mathbb{E}[N_2(T)]}{log(T)} \geq \frac{(1-\epsilon)}{(1+\epsilon)}\frac{1}{D(\mu_1, \mu_2)}$$

With respect to the bandit instance 1, the arm 2 is the suboptimal arm. So, from the equation above we see that, in order to find the optimal arm, number of plays of the suboptimal arm is proportional to $log(T)$.

This concludes the proof.

■

We saw in the proof above that for the game with 2 arms, the number of plays of suboptimal arm is lower bounded by order of $log(T)$. This can be extended to a game with K arms where $\forall i \neq i^*$, $\Delta_i > 0$, the number of times arm i is picked has a lower bound of the order $log(T)$.

References

[1] Bubeck, Sbastien, and Nicolo Cesa-Bianchi. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems." Foundations and Trends in Machine Learning 5.1 (2012): 1-122.