EE 290 Theory of Multi-armed Bandits and Reinforcement Learning	Lecture 5 - 2/2/2021

Lecture 5: Minimax Lower Bound for Finite-Arm Bandit Algorithms

Lecturer: Jiantao Jiao *Scribe: Elizabeth Glista, Kevin Lu, Nived Rajaraman*

In this lecture, we present an information-theoretic lower bound for the finite-arm i.i.d. bandits setting.

1 Information-theoretic lower bound for finite-arm i.i.d. bandits

Recall our previous setting: we have K arms and are playing for a horizon of T rounds with rewards sampled from $[0, 1]$. Our previously analysis for the ETC and UCB algorithms showed that they yield the upper bound $\mathbb{E}[R(T)] \lesssim \sqrt{KT \log T}$, where $R(T)$ is the pseudoregret. Today, we would like to justify our previous algorithms by showing that this upper bound is close to the information-theoretic lower bound. This lower bound is valid for any algorithm, i.e. it is a fundamental limit.

Previously, we considered the family of instances of the form $\nu \triangleq \{p_a : a \in \mathcal{A}\}$ where each p_a was a probability measure supported on the finite interval $[0, 1]$. Today, we consider the Gaussian family which has instances of the form $\nu \triangleq \{p_a = \mathcal{N}(\mu_a, 1) : a \in \mathcal{A}, 0 \le \mu_a \le 1\}$. Gaussians are easy to analyze and have nice properties such as exponential concentration inequalities. Even for this restricted family, we will see a lower bound of $\sqrt{KT}$ for the worst-case regret, implying that for a broader class of families we still cannot do better.[1]

In order to formulate this lower bound on worst-case regret, we will first introduce the notion of divergence.

Definition 1 (Kullback-Leibler (KL) divergence)**.** *For two probability measures P, Q on the same probability space, the KL divergence is defined as:*

$$D(P\|Q) \triangleq \begin{cases} \mathbb{E}_p\left[\log \frac{dP}{dQ}\right] & \text{if } P \ll Q \\ \infty & \text{otherwise} \end{cases} \tag{1}$$

where $\frac{dP}{dQ}$ is the likelihood ratio and $P \ll Q$ means that P is absolutely continuous w.r.t. Q, which is true if for any set A we have $Q(A) = 0 \Rightarrow P(A) = 0$.

Note that the condition $P \ll Q$ is critical in order for the ratio dP/dQ to be well-defined. When the probability measures P and Q have associated probability density functions, given by $p(x)$ and $q(x)$ respectively, then $\frac{dP}{dQ}(x) = \frac{p(x)}{q(x)}$, and we can write the KL divergence as $\int p(x) \log \frac{p(x)}{q(x)} dx$. For all probability measures P and Q, two important properties are that $D(P\|Q) \ge 0$ and $D(P\|Q) = 0 \iff P = Q$.

Example 2 (Likelihood ratio)**.** If we take Q to be the Lebesgue measure, i.e. $Q([0\ \ x]) = x$, and $P([0\ \ x]) = \int_0^x p(t)dt$, then the likelihood ratio is given by:

$$\frac{dP}{dQ} = \frac{d \int_0^x p(t)dt}{dx} = p(x)$$

Lemma 3 (Divergence Decomposition Lemma)**.** *Let $\nu = (p_1, p_2, \ldots, p_k)$ be one instance of rewards distributions for a bandit scenario, and $\nu' = (p'_1, p'_2, \ldots, p'_k)$ be another. Fix an arbitrary policy π consisting of the time-dependent policies $\pi_t(a_t|a_1, x_1, a_2, x_2, \ldots, a_{t-1}, x_{t-1})$ for $1 \le t \le T$. Let P_ν be the joint measure*

[1] Actually, using a KL divergence argument, we can derive the same result for other probability measures up to constants, so this lower bound holds for bounded random variables in general.

of $(A_1, X_1, A_2, X_2, \ldots, A_T, X_T)$ under instance ν and policy π, and $P_{\nu'}$ be defined similarity for instance ν' and policy π. Then the KL divergence between P_ν and $P_{\nu'}$ can be written as:

$$D(P_\nu \| P_{\nu'}) = \sum_{i=1}^{k} \mathbb{E}_\nu[n_T(i)] D(p_i || p_i') \tag{2}$$

where $n_T(i)$ is the number of times arm i was pulled by time T. Note that $n_T(i)$ is a random variable depending on both the randomness of the environment and the policy.

Proof First, note that we are defining $\frac{dP_\nu}{dP_{\nu'}}$ on the inputs $(A_1, X_1, A_2, X_2, \ldots, A_T, X_T)$. For any fixed sequence $(a_1, x_1, a_2, x_2, \ldots, a_T, x_T)$, we can write the joint distributions as follows:

$$P_\nu(a_1, x_2, a_2, x_2, \ldots, a_T, x_T) = \prod_{t=1}^{T} \pi_t(a_t | a_1, x_1, a_2, x_2, \ldots, a_{t-1}, x_{t-1}) p_{a_t}(x_t) \tag{3}$$

$$P_{\nu'}(a_1, x_2, a_2, x_2, \ldots, a_T, x_T) = \prod_{t=1}^{T} \pi_t(a_t | a_1, x_1, a_2, x_2, \ldots, a_{t-1}, x_{t-1}) p'_{a_t}(x_t) \tag{4}$$

Because the policy π is fixed in both P_ν and $P_{\nu'}$, when we consider $dP_\nu / dP_{\nu'}$, the π_t terms cancel out, leaving only the p_{a_t} and p'_{a_t} terms. This also removes the conditioning on the past history, simplifying the expectation. Hence, we can write the log-likelihood ratio cleanly as:

$$\log \frac{dP_\nu}{dP_{\nu'}} = \sum_{t=1}^{T} \log \frac{p_{a_t}(x_t)}{p'_{a_t}(x_t)} \tag{5}$$

Using this and the Law of Iterated Expectation, we can derive the result:

$$
\begin{aligned}
D(P_\nu \| P_{\nu'}) &= \mathbb{E}_\nu \left[\log \frac{dP_\nu}{dP_{\nu'}} \right] && (6) \\
&= \sum_{t=1}^{T} \mathbb{E}_\nu \left[\log \frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)} \right] && (7) \\
&= \sum_{t=1}^{T} \mathbb{E}_\nu \left[\mathbb{E} \left[\log \frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)} \Big| A_t \right] \right] && (8) \\
&= \sum_{t=1}^{T} \mathbb{E}_\nu \left[D(p_{A_t} || p'_{A_t}) \right] && (9) \\
&= \sum_{i=1}^{k} \mathbb{E}_\nu \left[\sum_{t=1}^{T} \mathbb{1}(A_t = i) D(p_{A_t} \| p'_{A_t}) \right] && (10) \\
&= \sum_{i=1}^{k} \mathbb{E}_\nu[n_T(i)] D(p_i || p_i') && (11)
\end{aligned}
$$

where we have introduced the indicator function $\mathbb{1}(\cdot)$ in Equation (10) and note that $n_T(i) = \sum_{t=1}^{T} \mathbb{1}(A_t = i)$ to yield Equation (11). □

One interpretation of this result is that given a particular KL divergence $D(P_\nu \| P_{\nu'})$ for an algorithm, if $D(p_i \| p_i')$ is small, you expect to have to pull arm i many times to figure out which instance you are in, while if $D(p_i \| p_i')$ is large, a good algorithm should be able to make the distinction with only a few pulls. Thus, the metric of $D(P_\nu \| P_{\nu'})$ is important for understanding how well an algorithm behaves.

Theorem 4. *For $T \geq K - 1$ and ν from the family of Gaussian bandit instances,*

$$\inf_{\pi} \sup_{\nu} \mathbb{E}[R(T)] \gtrsim \sqrt{KT} \tag{12}$$

Proof Let Δ be some real number in $\left[0, \frac{1}{2}\right]$. Choose mean vector μ in environment ν to be $(\Delta, 0, 0, \ldots, 0)$. Fix the policy π and compute $\mathbb{E}_{\nu}[n_T(i)]$ for each i.

Then, let $i = \arg\min_{j>1} \mathbb{E}_{\nu}[n_T(j)]$, the arm that is pulled the fewest number of times in expectation. Because $\sum_{j=1}^{k} \mathbb{E}_{\nu}[n_T(j)] = T$ and arm i is the least explored, we must have $\mathbb{E}_{\nu}[n_T(i)] \leq \frac{T}{k-1}$.

Now for environment ν', pick a new mean vector $\mu' = (\Delta, 0, \ldots, 0, 2\Delta, 0, \ldots, 0)$ where 2Δ is the reward on the i-th arm. Intuitively, we want to adversarially place a high reward on the arm that π pulls the least.

Define $\mathcal{R}_{\nu} \triangleq \mathbb{E}_{\nu}[R(T)]$ and $\mathcal{R}_{\nu'} \triangleq \mathbb{E}_{\nu'}[R(T)]$. The following inequalities follow from how often we pull arm 1, which has mean Δ:

$$\mathcal{R}_{\nu} \geq P_{\nu}\left(n_T(1) \leq \frac{T}{2}\right) \frac{T\Delta}{2} \tag{13}$$

$$\mathcal{R}_{\nu'} > P_{\nu'}\left(n_T(1) > \frac{T}{2}\right) \frac{T\Delta}{2} \tag{14}$$

where we have noted that in the first instance ν, the optimal strategy is to only pull arm 1 and in the second instance ν', the optimal strategy is to only pull the i-th arm. In the first instance ν, the suboptimality of choosing arm 1 only $T/2$ times is $T\Delta/2$, and in the second instance ν', the suboptimality of choosing arm 1 more than $T/2$ times is $T\Delta/2$. $\square$