CS861: Theoretical Foundations of Machine Learning Lecture 19 - 18/10/2023

University of Wisconsin–Madison, Fall 2023

Lecture 19: … lower bounds, generalized linear bandits

Lecturer: Kirthe… *Scribed by: Alex Clinton, Yamin Zhou*

Disclaimer: *These n… …jected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

In the previous lecture we proved an upper bound for the UCB and began analyzing a lower bound. In this lecture, we will continue the proof of that lower bound which is gap independent. We will then also provide a gap dependent lower bound for k-armed bandits. Finally, we will introduce a structured bandit model which generalizes the K–armed setting.

1 K-armed bandits lower bounds

Lemma 1. *Let ν, ν' be two bandit models, and P, P' bet the prob distribution of $A_1, x_1, ..., A_i, x_i$ due to the interaction of a policy π with ν, ν' respectively, then*

$$KL(P, P') = \sum_{i=1}^{k} \mathbb{E}_P[N_{i,T}] KL(\nu_i, \nu_i')$$

Proof we showed from the previous lecture, for any sequence $a_1, x_1, ...a_T, x_T$

$$log\left(\frac{p(a_1, x_1, ...a_T, x_T)}{p'(a_1, x_1, ...a_T, x_T)}\right) = \sum_{t=1}^{T} log\left(\frac{\hat{\nu}_{at}(x_t)}{\hat{\nu'}_{at}(x_t)}\right)$$

Therefore, we have

$$\begin{aligned}
KL(P, P') &= \mathbb{E}_p\left[\log\left(\frac{P(A_1, x_1.....A_i, x_i)}{P'(A_1, x_1......A_T, x_T)}\right)\right] \\
&= \mathbb{E}_p\left[\sum_{t=1}^{T} \log\left(\frac{\hat{\nu}_{At}(x_t)}{\hat{\nu'}_{At}(x_t)}\right)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_p\left[\log\left(\frac{\nu_{At}(x_t)}{\nu'_{At}(x_t)}\right) \sum_{i=1}^{K} \mathbb{1}(A_t = i)\right] \\
&= \sum_{i=1}^{K} \sum_{t=1}^{T} \mathbb{E}_p\left[\underbrace{\mathbb{E}_p\left[\log\left(\frac{\nu_{At}(x_t)}{\nu'_{At}(x_t)}\right) \mathbb{1}(A_t = i) | A_t\right]}_{*}\right] \\
&= \sum_{i=1}^{K} KL(\nu_i, \nu_i') \left(\sum_{t=1}^{T} \mathbb{E}[\mathbb{1}(A_t = i)]\right) \quad \textit{from the result of * and KL is not related to t} \\
&= \sum_{i=1}^{K} KL(\nu_i, \nu_i') \, \mathbb{E}[N_{i,T}]
\end{aligned}$$

where

$$i)\, \mathbb{E}_p \left[\log \left(\frac{\nu_{At}(x_t)}{\nu'_{At}(x_t)} \right) | A_t \right]$$
$$i) KL(\nu_{At}, \nu'_{At})$$
$$i) KL(\nu_i, \nu'_i)$$

□

This KL divergence decomposition lemma now allows us to state the following minimax lower bound for k-armed bandits. The idea of the proof is that given any policy π, we can construct two bandit instances such that π will perform poorly in one of them.

Theorem 1. *(Minimax lower bound for k-armed bandits) Let $P = \{\nu = \nu_i, i \in [K]\}, \nu_i$ is σ subGaussian for all $i \in [K]$. Then, if $K > 1$, for some universal constant C,*

$$\inf_{\pi} \sup_{\nu \in P} R_T(\pi, \nu) \geq C\sigma\sqrt{T(K-1)}$$

Proof Let π be given, consider the following two bandit models ν, ν' (constructed based on π) as follows:

- Let $\nu = \{\nu_i = N(\mu_i, \sigma^2)\}_{i \in [k]}$ where $\mu_1 = \delta$, and $\mu_i = 0 \; \forall i \in 2, \ldots, K$, $\mu = (\delta, 0, 0.....0)\}$. Here, $\delta > 0$ is a value we will specify shortly.
- Let $\mathbb{E}_\nu$ denote the expectation with respect to the sequence $A_1, x_1, ...A_T, x_T$ due to π's interaction with ν. Since $\sum_{i=1}^k \mathbb{E}_\nu[N_{i,t}] = T$, $\exists$ some $j \in 2, \ldots, k$ s.t $\mathbb{E}_\nu[N_{j,t}] \leq \frac{T}{k-1}$

- Let $\nu' = \{\nu_i = N(\mu'_i, \sigma^2)\}_{i \in [k]}$ where

$$\mu'_i = \begin{cases} \mu_i & if \quad i \neq j \\ 2\delta & if \quad i = j \end{cases}$$

  Therefore, $\mu' = (\delta, 0, 0, ..., \underbrace{2\delta}_{j}, 0, ..., 0)$

- Let P, P' denote the prob distributions of $A_1, x_1, ...A_T, x_T$ due to π's interaction with ν, ν'.

We know,

$$R_T(\pi, \nu) \geq P\left(N_{1,T} \leq \frac{T}{2}\right) \frac{T\delta}{2}$$
$$R_T(\pi, \nu') \geq P'\left(N_{j,T} \leq \frac{T}{2}\right) \frac{T\delta}{2} \geq P'\left(N_{1,T} \geq \frac{T}{2}\right) \frac{T\delta}{2}$$

Now, note that we can write

$$\sup_{\nu \in P} R_T(\pi, \nu) \geq \max(R_T(\nu, \pi), R_T(\nu', \pi)) \geq \frac{1}{2} \underbrace{(R_T(\nu, \pi) + R_T(\nu', \pi))}_{(\star)}$$

We will now bound the term $(\star)$ as follows

$$(\star) \geq \frac{T\delta}{2} \ldots {}_{1,T} > \frac{T}{2})\Big) \qquad \text{from the definition}$$
$$\geq \frac{T\delta}{4} \ldots \text{m the Le Cam's lemma}$$
$$= \frac{T\delta}{4} \ldots \qquad \text{Divergece Decomposition + KL of Gaussian}$$
$$\geq \frac{T\delta}{4} \ldots$$

Now, if we choose $\delta = \sigma\sqrt{\frac{K-1}{T}}$, Then, we are able to get $* \geq (\frac{1}{4}e^{-2})\sqrt{T(K-1)}$, and hence $\inf_\pi \sup_{\nu \in P} R_T(\pi, \nu) \geq C\sigma\sqrt{T(K-1)}$. $\square$

2 Gap-independent lower bounds

Recall from our analysis of UCB that there are two types of bounds: gap dependent bounds (those that depend on Δ_i) and gap independent bounds (those that do not depend on Δ_i). In addition to the gap independent lower bound we just proved, there is also a gap dependent lower bound for k-armed bandits. Although we will not prove it here, this lower bound is given by the following theorem.

Theorem 2. *(Theorem 16.4 in LS)*
Let ν be a given K-armed bandit model with σ-sub Gaussian rewards. Let $\mu = \mu(\nu)$ be the means of the arms. Let $\mathcal{P}(\nu) = \{\nu' : \mu_i(\nu) \in [\mu_i, \mu_i + 2\Delta_i], \nu_i' \text{ is } \sigma\text{-sub Gaussian}\}$. Say π is a policy such that $R_T(\pi, \nu') \leq cT^p, \forall \nu' \in \mathcal{P}(\nu)$ for some $c > 0$ and $p \in (0, 1)$. Then

$$R_T(\pi, \nu) \geq \frac{1}{2} \sum_{i:\Delta_i > 0} \frac{\left((1-p)\log(T) + \log(\frac{\Delta_i}{8c})\right)\sigma^2}{\Delta_i}$$

Remark At a high level this theorem says that if a policy does well on all "similar" problems then it does at least as poorly as the given expression on the original problem.

3 Stochastic bandits in a generalized linear model

One potential criticism of the bandit model we have studied thus far is its restriction of the action space to K specific choices. In the generalized linear bandit model we are about to introduce, we will allow for an infinite action space, but assume additional structure on the rewards.

Definition 1. *A generalized linear bandit model consists of the following components:*

1. Action space $\mathcal{A} \subseteq [-1, 1]^d$. For reasons of convenience which will become clear shortly, we will assume that the basis vectors $e_1, \ldots, e_d$ are in $\mathcal{A}$.
2. Parameter space $\Theta \subseteq [-1, 1]^d$
3. True parameter (unknown) $\theta_* \in \Theta$
4. When we choose an action (arm) A_t, we observe $X_t = f(\theta_*^T A_t) + \varepsilon_t$ where $\mathbb{E}[\varepsilon_t] = 0$ and ε_t is σ-sub Gaussian. Here ε_t can be thought of as noise.

5. Here f is known and has the following properties
   (a) f is strictly i[illegible] $c > 0$
   (b) f is L-Lipsch[illegible]
   (c) f' is continu[illegible]

Example (f is the ide[illegible]t model): In this case we have the following expression for the regret, $R_T = Tf(\theta_*^T$[illegible]here $a_* = \arg\max_{a \in \mathcal{A}} f(\theta_*^T a)$. Notice that this allows us to recover our original ban[illegible]w framework is broader than how we first introduced k-armed bandits.

3.1 A UCB algorithm (Based on Filippi et al. 2010)

In order to get a UCB algorithm we need to know how to estimate θ from data and construct UCBs. We define the following quantities:

- $\hat{\theta}_t = \arg\min_{\theta \in \Theta} \left\| \sum_{s=1}^{t} A_s (f(A_s^T \theta) - X_s) \right\|_{V_t^{-1}}$ where $v_t = \sum_{s=1}^{t} A_s A_s^T$ and $\|y\|_Q^2 = y^T Q y$ (here Q must be positive semi-definite)
- $\rho(t) = \frac{2L\sigma}{c} \sqrt{(3 + 2\log(1+2d)) \cdot 2d \log(t) \log(dT^2)} \in \tilde{O}(\sqrt{d})$

In the following algorithm, for the first d rounds, we pull each basis vector which is analogous to pulling each of the k arms once at the start of our original UCB algorithm. For the remaining rounds we then pull the arm with the highest new confidence bound which is again analogous the original UCB algorithm where we pull the arm with the highest original confidence bound.

Algorithm 1 UCB

Require: a time horizon T
  for $t = 1, \ldots, d$ **do**
    Choose $A_t = e_t$ (the t^{th} basis vector)
  end for
  for $t = d+1, \ldots, T$ **do**
    Choose $A_t = \arg\max_{a \in \mathcal{A}} \underbrace{f(\hat{\theta}_t^T a)}_{\text{exploitation}} + \underbrace{\rho(t)\|a\|_{V_{t-1}^{-1}}}_{\text{exploration}}$
  end for