CS861: Theoretical Foundations of Machine Learning Lecture 18 - 10/16/2023

University of Wisconsin–Madison, Fall 2023

Lecture 18: [illegible]cont'd), K-armed Bandit Lower Bound

Lecturer: Kirthe[illegible] *Scribed by: Michael Harding and Congwei Yang*

In this lecture, we will first upper bound the regret for UCB, providing gap-dependent and worst-case bounds. We will then start our discussion on proving lower bounds for K–armed bandits.

1 UCB Theorem and Proof

Recall the UCB algorithm from the last class.

Algorithm 1 The Upper Confidence Bound Algorithm

Require: time horizon T

for $t = 1, \ldots, K$ **do**

  $A_t \leftarrow t$

  $X_t \sim \nu_t$

end for

for $t = K+1, \ldots, T$ **do**

  $A_t \leftarrow \arg\max_{i\in[K]} \left(\hat{\mu}_{i,t-1} + e_{i,t-1}\right)$ ▷ Break ties arbitrarily

  $X_t \sim \nu_{A_t}$

end for

We will now present the theorem for the risk upper bounds for the UCB theorem once again, and pick up the proof where we left off.

Theorem 1 (UCB Risk Upper Bound). *Let* $\mathcal{P} = \left\{\nu = \{\nu_i\}_{i=1}^K : \nu_i\ \sigma\text{-}sG,\ \mathbb{E}_{X\sim\nu_i}[X] \in [0,1]\ \forall\, i \in [K]\right\}$ *be the class of* σ*-sub-Gaussian* K*-armed bandit models with means in* $[0,1]$*. Let* $\mu_i := \mathbb{E}_{X\sim\nu_i}[X]$*,* $\mu_* := \max_{i\in[K]} \mu_i$*, and denote* $\Delta_i := \mu_* - \mu_i$*. Then*

$$R_T(\nu) \leq 3K + \sum_{i:\Delta_i>0} \frac{24\sigma^2 \log(T)}{\Delta_i} \tag{1}$$

$$\sup_{\nu\in\mathcal{P}} R_T(\nu) \leq 3K + \sigma\sqrt{96KT\log(T)} \tag{2}$$

Proof As before, WLOG, we begin by letting $1 \geq \mu_1 \geq \cdots \geq \mu_K \geq 0$ for ease of notation. Also, we again define our good events

$$G_1 := \bigcap_{t>K} \left\{\mu_1 < \hat{\mu}_{1,t} + e_{1,t}\right\}$$

$$G_i := \bigcap_{t>K} \left\{\mu_i > \hat{\mu}_{i,t} - e_{i,t}\right\}$$

At the end of our previous class, we proved that $\mathbb{P}(G_1^c), \mathbb{P}(G_i^c) \leq \frac{1}{T}$ (we directly showed this for the case of G_1^c, remarking that the case for G_i^c is nearly identical). We will now show that $N_{i,t} \coloneqq \sum_{s=1}^{t} \mathbb{I}_{\{A_s = i\}}$ is small for sub-optimal arms ([illegible] t $G_1 \cap G_i$. To show this, suppose arm i was last pulled on round $t+1$, where $t \geq$ [illegible]

$$[illegible] + e_{j,t}) \leftarrow \text{UCB Alg. construction}$$

$$[illegible]\, t$$

$$[illegible] \text{r } G_1),$$

and under G_i, we also [illegible] erefore,

$$\mu_1 < \mu_i + 2e_{i,t} \Rightarrow \frac{\Delta_i}{2} < e_{i,t} = \sigma\sqrt{\frac{2\log(T^2 t)}{N_{i,t}}}$$

$$\Rightarrow N_{i,t} < \frac{8\sigma^2\log(T^3)}{\Delta_i^2} \leftarrow T > t$$

$$\Rightarrow N_{i,T} = N_{i,t} + 1 \leq \frac{24\sigma^2\log(T)}{\Delta_i^2} + 1$$

Now, combining these results, we can write,

$$\mathbb{E}[N_{i,t}] = \underbrace{\mathbb{E}[N_{i,t}|G_1 \cap G_i]}_{\leq \frac{24\sigma^2\log(T)}{\Delta_i^2}+1}\underbrace{\mathbb{P}(G_1 \cap G_i)}_{\leq 1} + \underbrace{\mathbb{E}[N_{i,t}|G_1^c \cup G_i^c]}_{\leq T}\underbrace{\mathbb{P}(G_1^c \cup G_i^c)}_{\leq \frac{2}{T}} \leq 3 + \frac{24\sigma^2\log(T)}{\Delta_i^2}$$

Then, by the regret decomposition result shown towards the end of last class, we can write,

$$R_T(\nu) \leq \sum_{i:\Delta_i > 0} \Delta_i\,\mathbb{E}[N_{i,t}] \leq 3K + \sum_{i:\Delta_i>0} \frac{24\sigma^2\log(T)}{\Delta_i},$$

where we leverage the fact that $\Delta_i \in [0,1]$ and there are at most $K-1$ summands. This proves the gap-dependent bound in (1). For the gap-independent bound, we can choose some value $\Delta > 0$ and rewrite our result above as thus:

$$\begin{aligned}
R_T(\nu) &= \sum_{i:\Delta_i>0} \Delta_i\,\mathbb{E}[N_{i,t}] \\
&= \sum_{i:\Delta_i \in (0,\Delta]} \Delta_i\,\mathbb{E}[N_{i,t}] + \sum_{i:\Delta_i > \Delta} \Delta_i\,\mathbb{E}[N_{i,t}] \\
&\leq \Delta \underbrace{\sum_{i:\Delta_i \in (0,\Delta]} \mathbb{E}[N_{i,t}]}_{\leq T} + \sum_{i:\Delta_i>\Delta} \frac{24\sigma^2\log(T)}{\Delta} + 3K \\
&\leq 3K + \Delta T + \frac{24\sigma^2\log(T)}{\Delta}
\end{aligned}$$

Then, because this holds for all $\Delta > 0$, we are free to optimize over values of Δ, giving us in particular $\Delta = \sigma\sqrt{\frac{24K\log(T)}{T}}$. Therefore,

$$R_T(\nu) \leq 3K + \sigma\sqrt{96KT\log(T)},$$

and because this result holds for all $\nu \in \mathcal{P}$, and the bound has no dependence on ν, then we can write,

$$\sup_{\nu\in\mathcal{P}} R_T(\nu) \leq 3K + \sigma\sqrt{96KT\log(T)},$$

which is exactly the statement in (2). $\square$

Next, we will presen[illegible] f the gap-independent bound. We will use similar techniques for linear bandits in sul[illegible]

1.1 Alternative [illegible]-Independent Bound

We will first decompose [illegible]

$$\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^{T}(\mu_* - X_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}[\mu_* - X_t \mid A_t]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}(\mu_* - \mu_{A_t})\right]
\end{aligned}$$

where $\mathbb{E}\left[\sum_{t=1}^{T}(\mu_* - \mu_{A_t})\right]$ is usually called the pseudo-regret. Let $G = G_1 \cap \bigcap_{i:\Delta_i>0} G_i$, then

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}(\mu_1 - \mu_{A_t}) \mid G\right] P(G) + \mathbb{E}\left[\sum_{t=1}^{T}(\mu_1 - \mu_{A_t}) \mid G^c\right] P(G^c) \tag{3}$$

Note we have $P(G) \leq 1$, $\mathbb{E}\left[\sum_{t=1}^{T}(\mu_1 - \mu_{A_t}) \mid G^c\right] \leq T$, and $P(G^c) \leq \frac{K}{T}$. We will bound $\sum_{t=1}^{T}(\mu_1 - \mu_{A_t})$ under G.

Claim: Under the event G, $\mu_1 - \mu_{A_t} \leq 2e_{A_t,t-1}$.

- If A_t is an optimal arm, then $\mu_1 - \mu_{A_t} \leq 0 \leq 2e_{A_t,t-1}$.
- If not, $\mu_1 \leq \hat{\mu}_{1,t-1} + e_{1,t-1} \leq \hat{mu}_{A_t,t-1} + e_{A_t,t-1} \leq \mu_{A_t} + 2e_{A_t,t-1}$, where the first inequality is under G_1, and the last inequality is under $\bigcap_{i:\Delta_i>0} G_i$.

Then,

$$\begin{aligned}
\sum_{t=1}^{T}(\mu_1 - \mu_{A_t}) &\leq K + \sum_{t=K+1}^{T} 2\sigma\sqrt{\frac{2\log(1/\delta_t)}{N_{A_t,t-1}}} \\
&\leq K + \sum_{t=K+1}^{T} 2\sigma\sqrt{\frac{2\log(T^2 t)}{N_{A_t,t-1}}} \\
&\leq K + \sigma\sqrt{24\log(T)} \sum_{t=K+1}^{T} \frac{1}{\sqrt{N_{A_t,t-1}}} \qquad (4)
\end{aligned}$$

We will now focus on the last summation:

$$
\begin{aligned}
\sum_{t=K\text{[illegible]}}^{T} \text{[illegible]} &\sum^{N_{i,T}-1}_{\text{[illegible]}} \frac{1}{\sqrt{s}} \\
&\text{[illegible]} \sqrt{N_{i,T}-1} \\
&\text{[illegible]} \left(\sum_{i=1}^{K} \sqrt{N_{i,T}-1} \right) \\
&\leq 2K \sqrt{\frac{1}{K} \sum_{i=1}^{K} (N_{i,T}-1)} \qquad (Jensen's\ Inequality) \\
&= 2\sqrt{K(T-K)} \qquad (5)
\end{aligned}
$$

Here the first inequality follows from $\sum_{s=1}^{m} \frac{1}{\sqrt{s}} \leq 2\sqrt{m}$, which we have proved below.

Combining (3), (4), (5), we obtain $R_T \leq 2K + \sigma\sqrt{96KT\log(T)}$. □

To prove, $\sum_{s=1}^{m} \frac{1}{\sqrt{s}} \leq 2\sqrt{m}$, we will bound the sum of a decreasing function by an integral as follows: $\sum_{s=1}^{m} \frac{1}{\sqrt{s}} \leq \int_0^m \frac{1}{\sqrt{s}}\, ds = (2s^{1/2})\big|_0^m = 2\sqrt{m}$.

2 K-armed bandits lower bound.

In this section, we will prove the following lower bound on the minimax regret: $\inf_{\Pi} \sup_{\nu \in \mathcal{P}} R_T(\Pi, \nu) \in \Omega(\sqrt{KT})$. To do so, recall the following results we used in the proof of Le Cam's method (Lecture 9, Lemma 1 and Corollary 1).

Lemma 1. *Let P_0, P_1 be two distributions and A be any event. Then,*

$$
\begin{aligned}
P_0(A) + P_1(A^c) &\geq \|P_0 \wedge P_1\| \qquad (Neyman-Pearson\ Test) \\
&= 1 - TV(P_0, P_1) \\
&\geq \frac{1}{2}\exp(-KL(P_0, P_1))
\end{aligned}
$$

When applying this inequality, the KL divergence will be between distributions of action-reward sequences $A_1, X_1, \cdots, A_T, X_T$ induced by the interaction of a policy π with different bandit models. The following lemma will be helpful in computing the KL divergence.

Lemma 2 (KL divergence decomposition)**.** *Let ν, ν' be two K-armed bandits models. For a fixed policy Π, let P, P' denote the probability distribution over the sequence of actions and rewards $A_1, X_1, \cdots, A_T, X_T$ under ν, ν', respectively. Let $\mathbb{E}_\nu$ denote the expectation under bandit model ν. Then $\forall T \geq 1$,*

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

where $N_{i,T} = \sum_{t=1}^{T} \mathbf{1}_{\{A_t = i\}}$

Intuitively, suppose we pulled arm 1 N_1 times. As the observations are independent $KL(P, P') = N_1 KL(\nu_1, \nu_1')$. Next, consider a nonadaptive policy which pulls arm i N_i times for $i = 1, \cdots, K$. We then have $KL(P, P') = \sum_{i=1}^{K} N_i KL(\nu_i, \nu_i')$. The above lemma says that a similar result holds when we use an adaptive policy.

Proof Proof of Lemma 2 Consider any given sequence $a_1, x_1, \cdots, a_T, x_T$. Let p, p' denote the Radon-Nikodym derivatives of P, P' respectively. Let $\tilde{\nu}_i, \tilde{\nu}'_i$ denote the Radon-Nikodym derivatives of ν_i, ν'_i, respectively.

Consider for fixed a $\cdots$ $_1, x_1, \cdots, a_T, x_T$.

$$p(a_1, x_1, \cdots, a_T, x_T) = \prod_{t=1}^{T} p(a_t, x_t \mid a_1, x_1, \cdots, a_{t-1}, x_{t-1})$$
$$= \prod_{t=1}^{T} \Pi(a_t \mid a_1, x_1, \cdots, a_{t-1}, x_{t-1}) \tilde{\nu}_{a_t}(x_t)$$

Similarly, under ν', we can write

$$p'(a_1, x_1, \cdots, a_T, x_T) = \prod_{t=1}^{T} \Pi(a_t \mid a_1, x_1, \cdots, a_{t-1}, x_{t-1}) \tilde{\nu}_{a_t}(x_t)$$

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

To be continued next lecture...

$\square$