CS861: Theoretical Foundations of Machine Learning Lecture 17 - 10/11/2024

University of Wisconsin–Madison, Fall 2024

Lecture 17: Proof for UCB (cont'd), K-armed Bandit Lower Bound

Lecturer: Kirthevasan Kandasamy *Scribed by: Guy Thampakkul and Haoqun Cao*

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

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 $\mu_i \coloneqq \mathbb{E}_{X\sim\nu_i}[X]$, $\mu_* \coloneqq \max_{i\in[K]} \mu_i$, and denote $\Delta_i \coloneqq \mu_* - \mu_i$. 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 Proof of Theorem 1 will assume w.l.o.g that each arm samples rewards $y_{i,r}{}_{r\in\mathbb{N}}$ and we observe these samples one-by-one as we pull each arm. Therefore, we can write $\hat{\mu}_{i,t} = \frac{1}{N_{i,t}} \sum_{r=1}^{N_{i,t}} y_{i,r}$.

Recall the definition of good events, G_1, G_i: for $\forall i$ s.t. $\Delta_i > 0$.

$$G_1 \triangleq \{\forall t > K,\ \mu_1 < \hat{\mu}_{1,t} + e_{1,t}\}$$

$$G_i \triangleq \{\forall t > K,\ \mu_i > \hat{\mu}_{i,t} - e_{i,t}\}$$

where G_1 indicates that the true mean is below the UCB, and G_i indicates that the true mean is above the LCB.

Claim 1. *We have, $\mathbb{P}(G_1^c) \leq \frac{1}{T}$, and $\mathbb{P}(G_i^c) \leq \frac{1}{T}$*

$$\begin{aligned}
\mathbb{P}(G_1^c) &= \mathbb{P}\left(\exists t > K,\ s.t.\ \mu_1 \geq \hat{\mu}_{1,t} + e_{1,t}\right) \\
&\leq \sum_{t>K} \mathbb{P}\left(\mu_1 > \hat{\mu}_{1,t} + e_{1,t}\right) \\
&= \sum_{t>K} \mathbb{P}\left(\mu_1 > \frac{1}{N_{1,t}} \sum_{r=1}^{N_{1,t}} y_{1,r} + \sigma\sqrt{\frac{2\log(1/\delta_t)}{N_{1,t}}}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{t>K} \mathbb{P}\left(\exists s \in [t-K+1] \text{ s.t. } \mu_1 > \frac{1}{s}\sum_{r=1}^{s} y_{1,r} + \sigma\sqrt{\frac{2\log(1/\delta_t)}{s}}\right) \\
&\leq \sum_{t>K} \sum_{s=1}^{t-K+1} \mathbb{P}\left(\frac{1}{s}\sum_{r=1}^{s}(y_{1,r}-\mu_1) < -\sigma\sqrt{\frac{2\log(1/\delta_t)}{s}}\right) \\
&\leq \sum_{t>K} \sum_{s=1}^{t-K+1} \exp\left(-\frac{s}{2\sigma^2}\cdot\sigma^2\cdot\frac{2\log(1/\delta_t)}{s}\right) \\
&= \sum_{t>K} \sum_{s=1}^{t-K+1} \frac{1}{T^2 t} \qquad \text{as } \delta_t = \frac{1}{T^2 t} \\
&\leq \sum_{t>K} \frac{1}{T^2} \leq \frac{1}{T}
\end{aligned}
$$

Remark The trick we used in the fourth and fifth steps only works in K–armed bandits. For other bandit models, we usually use martingales.

We will now show that $N_{i,t} \coloneqq \sum_{s=1}^{t} \mathbb{I}_{\{A_s=i\}}$ is small for sub-optimal arms ($\Delta_i > 0$) under the event $G_1 \cap G_i$. To show this, suppose arm i was last pulled on round $t+1$, where $t \geq K$. Hence,

$$
\begin{aligned}
\hat{\mu}_{i,t} + e_{i,t} &\geq \max_{j\neq i}\left(\hat{\mu}_{j,t} + e_{j,t}\right) \leftarrow \text{ UCB Alg. construction} \\
&\geq \hat{\mu}_{1,t} + e_{1,t} \\
&> \mu_1 \text{ (under } G_1\text{)},
\end{aligned}
$$

and under G_i, we also have $\mu_i > \hat{\mu}_{i,t} - e_{i,t}$. Therefore,

$$
\begin{aligned}
\mu_1 < \mu_i + 2e_{i,t} \Rightarrow \frac{\Delta_i}{2} < e_{i,t} &= \sigma\sqrt{\frac{2\log(T^2 t)}{N_{i,t}}} \\
&\Rightarrow N_{i,t} < \frac{8\sigma^2\log(T^3)}{\Delta_i^2} \leftarrow T > t \\
&\Rightarrow N_{i,T} = N_{i,t} + 1 \leq \frac{24\sigma^2\log(T)}{\Delta_i^2} + 1
\end{aligned}
$$

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). □

Next, we will present an alternative proof of the gap-independent bound. These techniques apply beyond K-armed bandits; we will use these techniques for linear bandits in subsequent classes.

1.1 Alternative Proof for the Gap-Independent Bound

We will first decompose the regret as follows:

$$\begin{aligned}
R_T &= T\mu_1 - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} (\mu_1 - X_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\mu_1 - X_t \mid A_t\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} (\mu_1 - \mu_{A_t})\right]
\end{aligned} \tag{3}$$

where $\mathbb{E}\left[\sum_{t=1}^{T} (\mu_1 - \mu_{A_t})\right]$ is usually called the pseudo-regret.

Next, we define the good event, $G = \bigcap_{i=1}^{K} G_i$, where

$$G_1 = \{\forall t > K, \mu_1 < \hat{\mu}_{1,t} + e_{1,t}\}, \quad G_i = \{\forall t > K, \mu_i > \hat{\mu}_{i,t} + e_{i,t}\}.$$

Note that above, $\forall T,\ \mathbb{P}(G_i^c) \leq \frac{1}{T}$.
Now, we rewrite R_T as follows

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \mu_1 - \mu_{A_t}\right] = \mathbb{E}\left[\sum_{t=1}^{T}(\mu_1 - \mu_{A_t}) \mid G\right] \cdot \underbrace{\mathbb{P}(G)}_{\leq 1} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}(\mu_1 - \mu_{A_t}) \mid G^c\right]}_{\leq T} \cdot \underbrace{\mathbb{P}(G^c)}_{\leq \frac{K}{T}} \tag{4}$$

We will bound $\sum_{t=1}^{T}(\mu_1 - \mu_{A_t})$ under G.
Claim: Under the event G, $\forall t > K$,

$$\mu_1 - \mu_{A_t} \leq 2e_{A_t,t-1} = 2\sigma\sqrt{\frac{2\log(T^2(t-1))}{N_{A_t,t-1}}}$$

.
Proof

$$\mu_1 \underbrace{\leq}_{\text{under } G} \hat{\mu}_{1,t-1} + e_{1,t-1} \underbrace{\leq}_{A_t \text{ was chosen on round } t} \hat{\mu}_{A_t,t-1} + e_{A_t,t-1} \underbrace{\leq}_{\text{under } G} \mu_{A_t} + 2e_{A_t,t-1}.$$

Therefore,

$$\mu_1 - \mu_{A_t} \leq 2e_{A_t,t-1} = 2\sigma\sqrt{\frac{2\log(T^2(t-1))}{N_{A_t,t-1}}}.$$

□

So, under G,

$$\sum_{t=1}^{T}(\mu_1 - \mu_{A_t}) = \underbrace{\sum_{t=1}^{K}(\mu_1 - \mu_{A_t})}_{K} + \underbrace{\sum_{t=K+1}^{T}(\mu_1 - \mu_{A_t})}_{\leq \sum_{t=K+1}^{T} 2\sigma\sqrt{\frac{2\log(T^3)}{N_{A_t,t-1}}}}$$

$$\leq K + \sigma\sqrt{24\log(T)} \underbrace{\sum_{t=K+1}^{T} \frac{1}{\sqrt{N_{A_t,t-1}}}}_{(*)} m$$

We proceed to bound **(*)** above

$$
\begin{aligned}
(*) &= \sum_{t=K+1}^{T} \frac{1}{\sqrt{N_{A_t,t-1}}} \\
&= \sum_{i=1}^{K} \sum_{s=1}^{N_{i,T}-1} \frac{1}{\sqrt{s}} \qquad \left(\sum_{s=1}^{m} \frac{1}{\sqrt{S}} \leq 2\sqrt{m}, \text{Proof below}\right) \\
&\leq 2 \sum_{i=1}^{K} \sqrt{N_{i,T}-1} \\
&= 2K \left(\frac{1}{K} \sum_{i=1}^{K} \sqrt{N_{i,T}-1}\right) \\
&\leq 2K \sqrt{\frac{1}{K} \underbrace{\sum_{i=1}^{K} (N_{i,T-1})}_{=T}} \qquad \text{(Jensen's Inequality)} \\
&= 2\sqrt{KT}
\end{aligned}
$$

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

$$
\sum_{t=1}^{T} \mu_1 - \mu_{A_t} \leq K + \sigma\sqrt{24\log(T)} \underbrace{\sum_{t=K+1}^{T} \frac{1}{\sqrt{N_{A_t,t-1}}}}_{(*) \leq \sqrt{2KT}} \leq K + \sigma\sqrt{96KT\log(T)}.
$$

Therefore,

$$
\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^{T} \mu_1 - \mu_{A_t} | G\right] \cdot \underbrace{\mathbb{P}(G)}_{\leq 1} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mu_1 - \mu_{A_t} | G^c\right]}_{\leq T} \cdot \underbrace{\mathbb{P}(G^c)}_{\leq K/T} \\
&\leq 2K + \sqrt{96KT\log(T)}.
\end{aligned}
$$

□

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:

$$
\inf_{\pi} \sup_{\nu \in \mathcal{P}} R_T(\pi, \nu) \in \Omega(?)
$$

.

For the UCB policy,

$$
R_T(\pi^{\text{UCB}}, \nu) \in \tilde{O}(\sqrt{KT}),\ \forall \nu \in \mathcal{P},
$$

where $\mathcal{P}$ is the set of $\sigma-$sub-Gaussian distributions.

We will prove a minimax lower bound of the above form through reduction to (binary) testing, i.e., we will do so by considering two alternatives and showing that no policy can simultaneously achieve small regret on both alternatives.

$$\inf_{\pi} \sup_{\nu \in \mathcal{P}} R_T(\pi, \nu) \geq \inf_{\pi} \sup_{\nu \in \{\nu_1, \nu_2\}} R_T(\pi, \nu) \geq \frac{1}{2}(R_T(\pi, \nu_1) + R_T(\pi, \nu_2)),$$

where the first inequality follows from the fact that the middle term is subset of the left term and the second inequality follows from the fact the max $\geq$ average.

To do so, recall the Bretagnolle-Huber inequality used in the proof of Le Cam's method.

Lemma 1. *Let P_0, P_1 be two distributions and A be any event. For any test ψ mapping the data to {0,1},*

$$P_0(\psi \neq 0) + P_1(\psi \neq 1) \underbrace{\geq}_{NP-test} ||P_0 \wedge P_1|| \underbrace{\geq}_{\text{by a propoerty we proved in class}} \frac{1}{2} e^{-KL(P_0, P_1)}.$$

We can write this as $\forall$ events $A, P_0(A) + P_1(A^c) \geq \frac{1}{2} e^{-KL(P_0, P_1)}$

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, which is not straightforward to compute. 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 given (possibly randomized) policy π, let P, P' denote the probability distribution over the sequence of actions and rewards $A_1, X_1, \ldots, A_T, X_T$ 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 $N_{i,T} = \sum_{t=1}^{T} \mathbb{1}(\{A_t = i\})$, and $\nu = \{\nu_i\}_{i \in [K]}, \nu' = \{\nu_i'\}_{i \in [K]}$.

Intuitively, suppose we pulled arm 1 N_1 times. As the observations are independent $KL(P, P') = N_1 KL(\nu_1, \nu_1')$.

To illustrate the intuition, let us consider a fixed policy which pulls arm i N_i times for $i = 1, \ldots, K$. We then have $KL(P, P') = \sum_{i=1}^{K} N_i KL(\nu_i, \nu_i')$.

$$\begin{aligned}
\mathrm{KL}(P, P') &= \mathbb{E}_\nu \left[\log \left(\frac{P(A_1, X_1, \ldots, A_T, X_T)}{P'(A_1, X_1, \ldots, A_T, X_T)} \right) \right] \\
&= \mathbb{E}_\nu \left[\log \left(\frac{P(\{\{Y_{i,r}\}_{r=1}^{N_i}\}_{i=1}^{K})}{P'(\{\{Y_{i,r}\}_{r=1}^{N_i}\}_{i=1}^{K})} \right) \right] \\
&= \mathbb{E}_\nu \left[\log \left(\frac{\prod_{i=1}^{K} \prod_{r=1}^{N_i} P_i(Y_{i,r})}{\prod_{i=1}^{K} \prod_{r=1}^{N_i} P_i'(Y_{i,r})} \right) \right] \\
&= \sum_{i=1}^{K} \mathbb{E}_\nu \left[\log \left(\frac{\prod_{r=1}^{N_i} P_i(Y_{i,r})}{\prod_{r=1}^{N_i} P_i'(Y_{i,r})} \right) \right]
\end{aligned}$$

$$= \sum_{i=1}^{K} \mathrm{KL}(\nu_i^{N_i}, \nu_i'^{N_i})$$
$$= \sum_{i=1}^{K} N_i \mathrm{KL}(\nu_i, \nu_i')$$

The divergence decomposition lemma says that a similar result holds when we use an adaptive policy, except with $N_{i,T}$ replaced with $\mathbb{E}[N_{i,T}]$. We will show the full proof of the above lemma next lecture.

To be continued next lecture...

Acknowledgement

These notes are based on scribed lecture materials prepared in Fall 2023 by Ransheng Guan, Yamin Zhou, Michael Harding, and Congwei Yang.