# Lecture 20: MAB Lower Bound

Course: Reinforcement Learning Theory
Instructor: Lei Ying
Department of EECS
University of Michigan, Ann Arbor

# Lower bound for MAB

## Kullback-Leibler Divergence (relative entropy)

For Bernoulli distributions $P, Q$ with parameters $p$ and $q$, their KL divergence is

$$kl(P,Q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$$

Consider two arms such that $\mu_1 > \mu_2$. Given $\epsilon > 0$, define $\mu_2'$ such that

(1) $\mu_1 < \mu_2' < 1$

(2) $kl(\mu_2, \mu_2') \leq (1+\epsilon) kl(\mu_2, \mu_1)$

- System 1: $\mu_1, \mu_2$
- System 2: $\mu_1, \mu_2'$

# Lower bound for MAB

System 1: $\mu_1, \mu_2$ System 2: $\mu_1, \mu_2'$

We want to show that:

(1) the forecaster cannot distinguish system 1 and 2 with a high probability, and

(2) the forecaster does not make too many mistakes in system 2

**Remark**

We can show (1) by lower bounding the number of times arm 2 is played in system 1, and (2) by lower bounding the number of times arm 2 is played in system 2.

# Lower bound for MAB

- Define $X_2(t)$: the reward when playing arm 2 the $t$-th time.
- Define

$$\hat{kl}_s = \sum_{t=1}^{s} \log \frac{\mu_2 X_2(t) + (1-\mu_2)(1-X_2(t))}{\mu_2' X_2(t) + (1-\mu_2')(1-X_2(t))}$$

- Then $\frac{\hat{kl}_s}{s}$ is an empirical estimate of $kl(\mu_2, \mu_2')$.

# Change-of-measure identity

$A$ : an event defined by $\{X_2(t)\}_{t=1,\ldots,T}$ and the MAB algorithm. Then,

$$\Pr'(A) = E\left[\mathbb{I}_A \exp(-\hat{k}l_{N_2(T)})\right]$$

Measured in system 2 Measured in system 2

# Example

Example: obtain reward 1 when playing arm 2 at the first time

$$\begin{aligned}
\Pr'(A) &= \Pr'(X_2(1)) = \mu_2' = \mu_2 \frac{\mu_2'}{\mu_2} \\
&= E\left[\mathbb{I}_{X_2(1)} \exp\left\{-\log \frac{\mu_2 X_2(1) + (1-\mu_2)(1-X_2(1))}{\mu_2' X_2(1) + (1-\mu_2')(1-X_2(1))}\right\}\right] \\
&= \mu_2 \times 1 \times e^{-\log \frac{\mu_2}{\mu_2'}} + (1-\mu_2) \times 0 \times e^{-\log \frac{1-\mu_2}{1-\mu_2'}} \\
&= \mu_2 \frac{\mu_2'}{\mu_2} = \mu_2'
\end{aligned}$$

■

# Lower bound for MAB

- We want to show that

$$\Pr\left(N_2(T) > \frac{1-\epsilon}{kl(\mu_2, \mu_2')} \log T\right) \approx 1$$

  i.e. arm 2 will be played at least $\log T$ times with a high probability, which leads to the $\log T$ lower bound.

- To prove this, we first show

$$\Pr\left(N_2(T) \leq \frac{1-\epsilon}{kl(\mu_2, \mu_2')} \log T \textbf{ and } \hat{kl}_{N_2(T)} \leq (1 - \frac{\epsilon}{2}) \log T\right) \approx 0, \text{i.e. } o(1)$$

# Lower bound for MAB

Intuition: Consider a "good" MAB algorithm such that for all $i$ such that $\Delta_i > 0$,

$$E[N_i(T)] = o(T^\alpha) \quad \forall 0 < \alpha \leq 1$$

In other words, a suboptimal arm is played no more than $T^\alpha$ times $\forall \alpha$. Otherwise, the regret is at least $\Delta_i T^\alpha > \log T$.

Thus,

- A "good" MAB algorithm should play arm 2 many times in system 2.
- If system 1 and system 2 are close from $\{X_{1,t}, X_{2,t}\}_{t=1,\ldots,T}$, then arm 2 should be played many times in system 1.

# Proof

- Define event

$$C_T = \left\{ N_2(T) < \frac{1-\epsilon}{kl(\mu_2, \mu_2')} \log T \text{ and } \hat{kl}_{N_2(T)} \leq (1 - \frac{\epsilon}{2}) \log T \right\}$$

$$\begin{aligned} \Pr'(C_T) &= E\left[ \mathbb{I}_{C_T} \exp(-\hat{kl}_{N_2(T)}) \right] \\ &\geq e^{-(1-\epsilon/2)\log T} \Pr(C_T) \end{aligned}$$

# Proof

Thus,

$$\begin{aligned}
\Pr(C_T) &\leq T^{1-\frac{\epsilon}{2}} P'(C_T) \\
&\leq T^{1-\frac{\epsilon}{2}} \Pr'(N_2(T) < \frac{1-\epsilon}{kl(\mu_2, \mu_2')} \log T) \\
&\leq T^{1-\frac{\epsilon}{2}} \frac{E'[T - N_2(T)]}{T - \frac{1-\epsilon}{kl(\mu_2, \mu_2') \log T}} \\
&\leq T^{1-\frac{\epsilon}{2}} \frac{T^{\alpha}}{T - \log T} \approx T^{-\frac{\epsilon}{4}} = o(1)
\end{aligned}$$

"Good" MAB

$\alpha = \frac{\epsilon}{4}$

# Proof

$$\widetilde{C}_T = \left\{ N_2(T) \leq \frac{1-\epsilon}{kl(\mu_2, \mu_2')} \log T \text{ and } \max_{s \leq \frac{1-\epsilon}{kl(\mu_2,\mu_2')} \log T} \hat{kl}_s \leq (1 - \frac{\epsilon}{2}) \log T \right\}$$

$$\widetilde{C}_T \subseteq C_T \implies \Pr(C_T) \geq \Pr(\widetilde{C}_T)$$

Note that

$$\max_{s \leq \frac{1-\epsilon}{kl(\mu_2,\mu_2')} \log T} \hat{kl}_s \leq (1 - \frac{\epsilon}{2}) \log T \quad \text{event B}$$

$$\Updownarrow$$

$$\frac{kl(\mu_2, \mu_2')}{(1-\epsilon)} \frac{1}{\log T} \max_{s \leq \frac{1-\epsilon}{kl(\mu_2,\mu_2')} \log T} \hat{kl}_s \leq \frac{1 - \frac{\epsilon}{2}}{1-\epsilon} kl(\mu_2, \mu_2')$$

# Proof

$$\frac{kl(\mu_2,\mu_2')}{(1-\epsilon)}\frac{1}{\log T}\max_{s\leq\frac{1-\epsilon}{kl(\mu_2,\mu_2')}\log T}\hat{kl}_s \leq \frac{1-\frac{\epsilon}{2}}{1-\epsilon}kl(\mu_2,\mu_2'),$$

where

$$\hat{kl}_s = \sum_{t=1}^{s}\log\frac{\mu_2 X_2(t)+(1-\mu_2)(1-X_2(t))}{\mu_2' X_2(t)+(1-\mu_2')(1-X_2(t))}$$

As $n\to\infty$,

$$\begin{aligned}E\left[\frac{1}{n}\hat{kl}_n\right] &= kl(\mu_2,\mu_2')\\ &< \frac{1-\frac{\epsilon}{2}}{1-\epsilon}kl(\mu_2,\mu_2').\end{aligned}$$

# Proof

By S.L.L.N for max,

$$\Pr(B) \to 1 \text{ as } T \to \infty$$

$$\implies \Pr\left(N_2(T) \leq \frac{1-\epsilon}{kl(\mu_2, \mu_2')} \log T\right) = o(1)$$

$$\implies R_T \geq \Delta_2 \frac{1}{1+\epsilon} \frac{1-\epsilon}{kl(\mu_2, \mu_1)} \log T$$

# Reference


- Chapter 2.3 of Bubeck, Sébastien, and Nicolo Cesa-Bianchi. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems." Foundations and Trends® in Machine Learning 5, no. 1 (2012): 1-122.



**Acknowledgements:** I would like to thank Alex Zhao for helping prepare the slides, and Honghao Wei and Zixian Yang for correcting typos/mistakes.