

Lecture 10: Stochastic Multi-Armed Bandits (Part II)

Week 10

Lecturer: Tianyu Wang

1 The UCB Algorithm

Consider the multi-armed problem with K arms. Let μ_i be the mean of the i -th distribution. Let $\hat{\mu}_{t,i}$ be the estimator of the mean of the i -th distribution at time t , which is defined as

$$\hat{\mu}_{t,i} = \frac{\sum_{s=1}^t Y_{I_s,s} \mathbb{I}_{[I_s=i]}}{\sum_{s=1}^t \mathbb{I}_{[I_s=i]}}.$$

Also define $n_{t,i} = \sum_{s=1}^t \mathbb{I}_{[I_s=i]}$.

At any $t \geq 1$, the UCB algorithm plays

$$I_t \in \arg \max_i \left\{ \hat{\mu}_{t,i} + \sqrt{\frac{6 \log t}{n_{t,i}}} \right\}.$$

1.1 Regret Analysis for UCB

Theorem 1.1 (Instance-dependent regret bound). *Given a **(fixed)** problem instance specified by K distributions supported on $[0, 1]$, the (expected) regret of the UCB algorithm (with $\delta = 1/T^2$) satisfies*

$$\mathbb{E} \left[\sum_{t=1}^T \mu_* - \sum_{t=1}^T \mu_{I_t} \right] \leq O \left(\sum_{k=1}^K \frac{\log T}{\Delta_k} + K \right), \quad \forall \text{ known constant } T \geq 1,$$

where $\Delta_k = \mu_* - \mu_i$ with μ_i being the expectation of the i -th distribution, and $\mu_* = \max_{i: 1 \leq i \leq K} \mu_i$.

Proof. Rewrite the regret as follows.

$$\begin{aligned}
\mathbb{E} \left[\sum_{t=1}^T \mu_* - \sum_{t=1}^T \mu_{I_t} \right] &= \mathbb{E} \left[\sum_{t=1}^T \sum_{k=1}^K \mu_* \mathbb{I}_{[I_t=k]} \right] - \mathbb{E} \left[\sum_{t=1}^T \sum_{k=1}^K \mu_{I_t} \mathbb{I}_{[I_t=k]} \right] \\
&= \sum_{k=1}^K \mu_* \mathbb{E} [n_{T,k}] - \sum_{k=1}^K \mu_{I_k} \mathbb{E} [n_{T,k}] \\
&= \sum_{k=1}^K (\mu_* - \mu_k) \mathbb{E} [n_{T,k}] \\
&= \sum_{k=1}^K \Delta_k \mathbb{E} [n_{T,k}],
\end{aligned}$$

where $\Delta_k := \mu_* - \mu_k$.

Recall last time we proved that, at any fixed δ , with probability greater than $1 - 2\delta$,

$$|\hat{\mu}_{t,i} - \mu_i| \leq \sqrt{\frac{2 \log(1/\delta)}{n_{t,i}}}, \quad \forall i = 1, 2, \dots, K.$$

Thus we have, with high probability greater than $1 - K\delta$,

$$\begin{aligned}
\Delta_{I_t} &= \mu_* - \mu_{I_t} \\
&\leq \hat{\mu}_{t,*} + \sqrt{\frac{2 \log(1/\delta)}{n_{t,*}}} - \hat{\mu}_{t,I_t} + \sqrt{\frac{2 \log(1/\delta)}{n_{t,*}}} && \text{(by Azuma-Hoeffding)} \\
&\leq \hat{\mu}_{t,I_t} + \sqrt{\frac{2 \log(1/\delta)}{n_{t,i}}} - \hat{\mu}_{t,I_t} + \sqrt{\frac{2 \log(1/\delta)}{n_{t,i}}} && \text{(since } I_t \in \arg \max_i \left\{ \hat{\mu}_{t,i} + \sqrt{\frac{6 \log t}{n_{t,i}}} \right\}) \\
&= 2\sqrt{\frac{2 \log(1/\delta)}{n_{t,I_t}}}, && (1)
\end{aligned}$$

which implies

$$n_{t,I_t} \leq \frac{8 \log(1/\delta)}{\Delta_{I_t}^2}$$

provided that $\Delta_{I_t} > 0$.

Let \mathcal{E}_t be the event:

$$\mathcal{E}_t = \left\{ |\hat{\mu}_{t,i} - \mu_i| \leq \sqrt{\frac{2 \log(1/\delta)}{n_{t,i}}}, \quad \forall i = 1, 2, \dots, K. \right\}.$$

If any of \mathcal{E}_t is violated, we have

$$\mathbb{E} \left[\sum_{t=1}^T (\mu_* - \mu_{I_t, t}) \mid \overline{\cap_{t=1}^T \mathcal{E}_t} \right] \mathbb{P} \left(\overline{\cap_{t=1}^T \mathcal{E}_t} \right) \leq T \cdot \delta K T. \quad (2)$$

Let τ_k^{last} be the last time k is played (conditioning on $\cap_{t=K+1}^\infty \mathcal{E}_t$). The regret satisfies

$$\begin{aligned} \sum_{k=1}^K \Delta_k \mathbb{E} [n_{T,k} \mid \cap_{t=1}^T \mathcal{E}_t] &= \sum_{k=1}^K \Delta_k \mathbb{E} [n_{\tau_k^{last}, k} \mid \cap_{t=K+1}^\infty \mathcal{E}_t] \\ &\leq \sum_{k=1}^K \frac{24 \log T}{\Delta_k}. \end{aligned}$$

Thus the regret satisfies

$$\mathbb{E} \left[\sum_{t=1}^T \mu_* - \sum_{t=1}^T \mu_{I_t} \right] \leq \sum_{k=1}^K \frac{24 \log T}{\Delta_k} + \delta K T^2.$$

Picking $\delta = \frac{1}{T^2}$ finished the proof. \square

Theorem 1.2 (Instance-independent regret bound). *For **any** problem instance specified by K distributions supported on $[0, 1]$, the (expected) regret of the UCB algorithm satisfies*

$$\mathbb{E} \left[\sum_{t=1}^T \mu_* - \sum_{t=1}^T \mu_{I_t} \right] \leq \mathcal{O} \left(\sqrt{KT \log(T)} + K \right)$$

for any (known) constant $T \geq 1$.

Proof. By (1), we have

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^T \mu_* - \mu_{I_t} \mid \cap_{t=K+1}^\infty \mathcal{E}_t \right] &\leq \mathbb{E} \left[\sum_{t=1}^T 2 \sqrt{\frac{6 \log t}{n_{t, I_t}}} \mid \cap_{t=K+1}^\infty \mathcal{E}_t \right] \\ &\leq \mathbb{E} \left[\sum_{i=1}^K \sum_{m=1}^{n_{T,i}} 2 \sqrt{\frac{6 \log T}{m}} \mid \cap_{t=K+1}^\infty \mathcal{E}_t \right] \quad (\text{regroup the terms}) \\ &\leq \mathbb{E} \left[\sum_{i=1}^K 2 \sqrt{18 n_{T,i} \log T} \mid \cap_{t=K+1}^\infty \mathcal{E}_t \right] \\ &\leq \mathbb{E} \left[\sqrt{K} \sqrt{\sum_{i=1}^K 36 n_{T,i} \log T} \mid \cap_{t=K+1}^\infty \mathcal{E}_t \right] \\ &\quad (\text{by Cauchy-Schwarz inequality}) \\ &\leq 6 \sqrt{KT \log T}. \quad (\text{since } \sum_{i=1}^K n_{T,i} = T) \end{aligned}$$

Combining the above result with (2) concludes the proof. \square

2 Lower Bounds

Theorem 2.1 (worst case lower bound). *Fix the number of distributions (arms) to K . For any fixed time horizon T , there exists a problem instance, such that the regret for all algorithms is of order $\Omega(\sqrt{KT})$.*

The above result is also known as minimax lower bound or instance-independent lower bound.

Definition 2.2. Consider a measurable space (X, Σ) . For two probability measures \mathbb{P} and \mathbb{Q} defined on the measurable space (X, Σ) , the total variation between \mathbb{P} and \mathbb{Q} is

$$\|\mathbb{P} - \mathbb{Q}\|_{TV} = 2 \sup\{|\mathbb{P}(A) - \mathbb{Q}(A)| : A \in \Sigma\}.$$

Theorem 2.3 (Pinsker's inequality). *For any two probability measures \mathbb{P} and \mathbb{Q} defined on the same measurable space (X, Σ) , it holds that*

$$\|P - Q\|_{TV} \leq \sqrt{2D_{KL}(P\|Q)}.$$

Proposition 2.4 (Chain rule for KL-divergence). *Let \mathbb{P} and \mathbb{Q} be two probability measures defined on the same space (X, Σ) , and let two random variables X and Y be measurable with respect to (X, Σ) . Then we have*

$$D_{KL}(\mathbb{P}(X, Y)\|\mathbb{Q}(X, Y)) = D_{KL}(\mathbb{P}(X)\|\mathbb{Q}(X)) + D_{KL}(\mathbb{P}(Y|X)\|\mathbb{Q}(Y|X)).$$

Recall the KL-divergence for two conditional distributions are

$$D_{KL}(\mathbb{P}(Y|X)\|\mathbb{Q}(Y|X)) = \sum_x \mathbb{P}(x) \sum_y \mathbb{P}(y|x) \log \frac{\mathbb{P}(y|x)}{\mathbb{Q}(y|x)}$$

The proof for the above proposition is similar to Q11 in Quiz 1.

Proof of Theorem 2.1. Construct $K+1$ Bernoulli instances (all distributions/arms are Bernoulli) as follows: in \mathfrak{J}_0 the means of the Bernoulli distributions are $(\frac{1}{2}, \frac{1}{2}, \dots, \frac{1}{2})$, $\mathfrak{J}_1 = (\frac{1}{2} + \epsilon, \frac{1}{2}, \frac{1}{2}, \dots, \frac{1}{2})$, $\mathfrak{J}_2 = (\frac{1}{2}, \frac{1}{2} + \epsilon, \frac{1}{2}, \dots, \frac{1}{2})$, \dots , $\mathfrak{J}_K = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \dots, \frac{1}{2} + \epsilon)$ for some ϵ to be specified later.

Step 1: compute the KL-divergence between \mathfrak{J}_0 and \mathfrak{J}_k .

For any policy π , let $\mathcal{P}_{k,\pi}$ be the probability measure of executing policy π on instance \mathfrak{J}_k .

Note that

$$\begin{aligned} D_{KL}\left(\text{Bernoulli}\left(\frac{1}{2}\right) \parallel \text{Bernoulli}\left(\frac{1}{2} + \epsilon\right)\right) &= \frac{1}{2} \log\left(\frac{1/2}{1/2 + \epsilon}\right) + \frac{1}{2} \log\left(\frac{1/2}{1/2 - \epsilon}\right) \\ &\geq 2\epsilon^2. \end{aligned}$$

By chain rule of KL-divergence, we have, for any $k = 1, \dots, K$,

$$\begin{aligned} D_{KL}(\mathbb{P}_{0,\pi} \parallel \mathbb{P}_{k,\pi}) &= \sum_{t=1}^T \sum_{j=1}^K \mathbb{P}_{0,\pi}(I_t = j) D_{KL}\left(\text{Bernoulli}\left(\frac{1}{2}\right) \parallel \text{Bernoulli}\left(\frac{1}{2} + \epsilon \mathbb{I}_{[I_t=j]}\right)\right) \\ &= 2\epsilon^2 \mathbb{E}_{0,\pi}[n_{T,k}]. \end{aligned} \tag{3}$$

Step 2: the optimality gap between arms.

In this case, the optimality gap is trivially ϵ .

Step 3: apply Yao's principle and Pinsker's inequality to finish the proof.

By Pinsker's inequality, we have $\forall j, k$,

$$|\mathbb{P}_{0,\pi}(I_t = j) - \mathbb{P}_{k,\pi}(I_t = j)| \leq \sqrt{2D_{KL}(\mathbb{P}_{0,\pi} \parallel \mathbb{P}_{k,\pi})}. \quad (4)$$

Thus for the regret against k is instance k , we have

$$\begin{aligned} & \max_{k \in [K]} \sum_{t=1}^T (\mathbb{E}_{k,\pi}[Y_{k,t}] - \mathbb{E}_{k,\pi}[Y_{I_t,t}]) \\ & \geq \frac{1}{K} \sum_{k=1}^K \sum_{t=1}^T \mathbb{E}_{k,\pi}[Y_{k,t}] - \mathbb{E}_{k,\pi}[Y_{I_t,t}] \\ & = \frac{\epsilon}{K} \sum_{k=1}^K \sum_{t=1}^T \mathbb{P}_{k,\pi}(I_t \neq k) \quad (\text{by the Wald's identity}) \\ & = \epsilon T - \frac{\epsilon}{K} \sum_{k=1}^K \sum_{t=1}^T \mathbb{P}_{k,\pi}(I_t = k) \\ & \geq \epsilon T - \frac{\epsilon}{K} \sum_{k=1}^K \sum_{t=1}^T \mathbb{P}_{0,\pi}(I_t = k) - \frac{\epsilon}{K} \sum_{k=1}^K \sum_{t=1}^T \sqrt{2D_{KL}(\mathcal{P}_{0,\pi} \parallel \mathcal{P}_{k,\pi})} \quad (\text{by Eq. 4}) \\ & \geq \frac{(K-1)\epsilon T}{K} - 2\epsilon^2 T \sqrt{\frac{1}{K} \sum_{k=1}^K 2\mathbb{E}_{0,\pi}[n_{T,k}]} \quad (\text{by Jensen's inequality}) \\ & = \frac{(K-1)\epsilon T}{K} - 2\epsilon^2 T \sqrt{\frac{T}{K}}. \end{aligned} \quad (5)$$

Since the above bound is true for any ϵ , letting $\epsilon = \sqrt{\frac{4K}{T}}$ concludes the proof. \square

Definition 2.5 (consistent policies). A policy π is consistent (over a set of problem instances) if for all problem instances (in the set) the regret incurred by policy π after T steps (denoted $R(\pi, T)$) satisfies

$$\lim_{T \rightarrow \infty} \frac{R(\pi, T)}{T^\alpha} \leq 1,$$

for all $\alpha > 0$.

Theorem 2.6 (asymptotic lower bound (for consistent policies)). *Given any Bernoulli problem instance $\mathfrak{J} = (\mu_1, \mu_2, \dots, \mu_K)$ and a consistent policy π , it holds that, for any suboptimal distribution/arm i ,*

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}_{\mathfrak{J},\pi}[n_{T,i}]}{\log T} \geq \frac{1}{D_{KL}(\text{Bernoulli}(\mu_i) \parallel \text{Bernoulli}(\mu_i + \Delta_i))},$$

where $\mathbb{E}_{\mathfrak{J},\pi}$ is the expectation with respect to the randomness generated by instance \mathfrak{J} and policy π .

Theorem 2.7 (Bretagnolle-Huber-Tsybakov inequality). *For any probability measures \mathbb{P}, \mathbb{Q} on (X, Σ) , it holds that*

$$\|\mathbb{P} - \mathbb{Q}\|_{TV} \leq 1 - \frac{1}{2} \exp(-D_{KL}(\mathbb{P} \parallel \mathbb{Q})).$$

Proof of Theorem 2.6. For the Bernoulli problem instance $\mathfrak{J} = (\mu_1, \mu_2, \dots, \mu_K)$, and let i be a sub-optimal arm in \mathfrak{J} . Let \mathfrak{J}' be another instance such that $\mathfrak{J}' = (\mu_1, \mu_2, \dots, \mu_{i-1}, \mu'_i, \mu_{i+1}, \dots, \mu_K)$, in where all distributions, except for the i -th one, are identical to those in \mathfrak{J} . The value of μ'_i will be specified later.

By chain rule of KL-divergence, it holds that

$$D_{KL}(\mathbb{P}_{\mathfrak{J},\pi} \parallel \mathbb{P}_{\mathfrak{J}',\pi}) = \mathbb{E}_{\mathfrak{J},\pi}[n_{T,i}] D_{KL}(\mu_i \parallel \mu'_i),$$

where $D_{KL}(\mu \parallel \mu')$ is a shorthand for $D_{KL}(\text{Bernoulli}(\mu) \parallel \text{Bernoulli}(\mu'))$, which is of order $O((\mu_i - \mu'_i)^2)$.

By the Bretagnolle-Huber-Tsybakov inequality, we have

$$\mathbb{P}_{\mathfrak{J},\pi}(\{n_{T,i} \geq T/2\}) - \mathbb{P}_{\mathfrak{J}',\pi}(\{n_{T,i} \geq T/2\}) \leq \|\mathbb{P}_{\mathfrak{J},\pi} - \mathbb{P}_{\mathfrak{J}',\pi}\| \leq 1 - \frac{1}{2} \exp(-D_{KL}(\mathbb{P}_{\mathfrak{J},\pi} \parallel \mathbb{P}_{\mathfrak{J}',\pi})).$$

Let $\mu'_i = \mu_i + \lambda$ where $\lambda > \Delta_i$. Let $R(\pi, T)$ (resp. $R'(\pi, T)$) be the expected first T step regret of π in \mathfrak{J} (resp. \mathfrak{J}').

By Markov inequality, we have

$$R(\pi, T) \geq \Delta_i \mathbb{E}_{\mathfrak{J},\pi}[n_{T,i}] \geq \frac{T\Delta_i}{2} \mathbb{P}_{\mathfrak{J},\pi}\left(n_{T,i} \geq \frac{T}{2}\right).$$

Also, by writing out the conditional expectation, we have

$$R'(\pi, T) \geq \frac{T(\lambda - \Delta_i)}{2} \mathbb{P}_{\mathfrak{J}',\pi}\left(n_{T,i} < \frac{T}{2}\right).$$

Thus we have

$$\begin{aligned} D_{KL}(\mu_i \parallel \mu'_i) \mathbb{E}_{\mathfrak{J},\pi}[n_{T,i}] &= D_{KL}(\mathbb{P}_{\mathfrak{J},\pi} \parallel \mathbb{P}_{\mathfrak{J}',\pi}) \\ &\geq \log(2\mathbb{P}_{\mathfrak{J},\pi}(\{n_{T,i} \geq T/2\}) + 2\mathbb{P}_{\mathfrak{J}',\pi}(\{n_{T,i} < T/2\})) \\ &\geq \log \frac{T \min\{\Delta_i, \lambda - \Delta_i\}}{4R(\pi, T) + 4R'(\pi, T)}. \end{aligned}$$

Thus we have

$$\frac{\mathbb{E}_{\mathfrak{J},\pi}[n_{T,i}] D_{KL}(\mu_i \parallel \mu_i + \lambda)}{\log T} \geq 1 + \frac{\log \min\{\Delta_i, \lambda - \Delta_i\}}{\log T} - \frac{\log(4R(\pi, T) + 4R'(\pi, T))}{\log T}.$$

Since the policy is consistent, we know that $\frac{\log(4R(\pi,T)+4R'(\pi,T))}{\log T} = \frac{\log(O(T^p))}{\log T} = p$ for any $p > 0$. Thus we have

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}_{\mathfrak{J}, \pi} [n_{T,i}] D_{KL}(\mu_i \| \mu_i + \lambda)}{\log T} \geq 1.$$

Since the above is true for all $\lambda > \Delta_i$ and the KL-divergence is continuous, we have

$$\inf_{\lambda > \Delta_i} D_{KL}(\mu_i \| \mu_i + \lambda) = D_{KL}(\mu_i \| \mu_i + \Delta_i).$$

Rearranging terms gives

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}_{\mathfrak{J}, \pi} [n_{T,i}]}{\log T} \geq \frac{1}{D_{KL}(\mu_i \| \mu_i + \Delta_i)}.$$

□

Acknowledgement

The modern form of stochastic MAB problem is due to Lai & Robbins. A recommended reference for this part is the Bandit Algorithms book by Lattimore and Szepesvari. The simple instance-dependent upper bound proof in the notes is provided by Feng.