Hellinger-UCB: A novel algorithm for stochastic multi-armed bandit problem

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we study the stochastic multi-armed bandit problem, where the reward is driven by an unknown random variable. We propose a new variant of the Upper Confidence Bound (UCB) algorithm called Hellinger-UCB, which leverages the squared Hellinger distance to build the upper confidence bound. We prove that the Hellinger-UCB reaches the theoretical lower bound(O(T)). As a real-world example, we apply the Hellinger-UCB algorithm to solve the cold-start problem for a content recommender system of a financial app. With reasonable assumption, the Hellinger-UCB algorithm has an important lower latency feature, closed-form UCB. The online experiment also illustrates that the Hellinger-UCB outperforms both KL-UCB and UCB1 in the sense of a higher click-through rate (CTR), 33% higher than the KL-UCB and almost 100% higher than the UCB1.

1 Introduction

1.1 Stochastic multi-armed bandit problem

The stochastic multi-armed bandit problem(MAB) [10] is a sequential decision problem defined by a payoff function and a set of actions. At each step $t \in 1, 2, ..., T$, an action A_t is chosen from the action set $A = \{1, 2, 3, ...K\}$ by the agent. And the associated reward $r_t(A_t)$, which is independent and identically distributed(i.i.d.), is obtained. The goal of the agent is to find the optimal strategy that maximizes the cumulative payoff obtained in a sequence of decisions

$$S_{A_t}(T) = \sum_{s=1}^{T} r_t(A_t). \tag{1}$$

The agent must come up with a strategy to maximize the cumulative payoff by dealing with the dilemma between exploitation and exploration. The pseudo-regret $\bar{R}_T$ is introduced to evaluate the performance of a strategy. It is defined as the maximized expectation of the difference between the cumulative payoff of consistently choosing the best action and that of the strategy in the first T steps

$$\bar{R}_T = \max_{i=1,2,...,K} \mathbb{E}[\sum_{s=1}^{T} r_t(i) - \sum_{s=1}^{T} r_t(A_t)] \tag{2}$$

Lai and Robbins(1985)[8] showed that if for all $\epsilon > 0$ and $\theta, \theta_i \in \Theta$ with $\mu(\theta) > \mu(\theta_i)$, there exists $\delta > 0$ such that $|KL(\theta, \theta_i) - KL(\theta, \theta_j)| < \epsilon$ whenever $\mu(\theta_i) < \mu(\theta_j) < \mu(\theta_i) + \delta$, and the following theorem is true. Let $N_i(t)$ denote the number of times the agent selected action i in the first t steps,

Submitted to Workshop on Bayesian Decision-making and Uncertainty, 38th Conference on Neural Information Processing Systems (BDU at NeurIPS 2024). Do not distribute.

Theorem 1 *If a policy has regret $\bar{R}_T = o(T^a)$ for all $a > 0$ as $T \to 0$, the number of draws up to time t, $N_i(t)$ of any sub-optimal arm i is lower bounded*

$$\lim \inf_{T\to\infty} \frac{N_i(T)}{\log(T)} \geqslant \frac{1}{\inf_{\theta\in\Theta_i:\mathbb{E}_\theta>\mu^*} KL(\theta_i, \theta)} \tag{3}$$

Therefore, the regret is lower-bounded

$$\lim \inf_{T\to\infty} \frac{\bar{R}_T}{\log(T)} \geqslant \sum_{i:\Delta_i\geqslant 0} \frac{\Delta_i}{\inf_{\theta\in\Theta_i:\mathbb{E}_\theta>\mu^*} KL(\theta_i, \theta)} \tag{4}$$

.

The stochastic multi-armed bandit problem has been extensively studied [4, 7, 9]. Under the parametric setting, a type of policy called upper-confidence bound (UCB) is proposed and proved to be promising[3]. Agrawal[1] introduced a family of index policies that is easier to compute. Auer, Nicolo and Paul[3] proposed an online, horizon-free procedure which is called upper-confidence bound(UCB) and proved its efficiency. Audibert and Bubeck[2] presented an improvement to the UCB1 called MOSS which is optimal for finite time. A variant of UCB which builds UCB based on the Kullbakc-Leibler divergence(KL) KL-UCB was presented by Garivier, Cappé[6].

2 Setup And Notations

We consider a stochastic multi-armed bandit problem with finite arms $A = \{1, 2, 3, ...K\}$. Each arm i is associated with a reward distribution $p_i(\theta)$ over $\mathbb{R}$. It is assumed that $p_i(\theta)$ is from some one-parameter exponential family $\mathbb{P}(\theta)$ with unknown expectation μ_i.

It is also common to use a different formula for the pseudo-regret for a stochastic problem. Write $\mu^* = \max_{i=1,2,...,K} \mathbb{E}[\mu_i]$ as the expected reward of the optimal action. Then $\Delta_i = \mu_* - \mu_i$, and we have

$$\bar{R}_T = \sum_{i=1}^{K} \mathbb{E}[N_i(T)]\mu^* - \mathbb{E}[\sum_{i=1}^{K} N_i(T)\mu_i] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[N_i(T)] \tag{5}$$

3 The Hellinger-UCB Algorithm

The goal of the UCB algorithm is to make sequential decisions in the stochastic environment. The reward distribution of each arm is unknown. The only way to collect information and estimate the distribution is to pull the arm. But each trial comes with risk which is measured by regret. Hence exploration-exploitation trade-off is important in this case. The motivation of the UCB algorithm is being optimistic about the reward distributions as one always believes that the expected reward is the highest value within the confidence region. Hence the key in the UCB algorithm is constructing the confidence region.

The formula of the squared Hellinger distance makes it computationally efficient. With this property, we propose a new UCB type algorithm, the Hellinger-UCB which constructs the UCB based on the squared Hellinger distance. The new algorithm achieves the theoretical lower bound and has a closed-form UCB for some distributions, for example, binomial distribution. The latter property is favorable in some low-latency applications.

3.1 Main algorithm

We briefly describe the process of Hellinger-UCB here. Let $A = \{i\}_{i=1}^{K}$ be the action set where K, the number of actions, is a positive integer. For each arm $i \in A$, the reward distribution P_{θ_i} is in some one-parameter exponential family with expectation μ_i. At the first $|K|$ rounds, the agent chooses each arm once. After that, at each round $t > |K|$, the agent makes a decision $A_t = i$ based on the collected observations of each arm and gets the reward $g_t(A_t)$ from P_{A_t}. The upper confidence bound for arm i is

$$U_i(t) = \sup\{\dot{\psi}(\theta) : H^2(P_{\hat{\theta}_{i,t-1}}, P_\theta) \leqslant 1 - e^{-c\frac{\log(t)}{N_i(t)}}\} \tag{6}$$

where $P_{\hat{\theta}_{i,t-1}}$ is the estimated reward distribution based on the past observations and $N_i(t)$ the number of pulls of arm i. In the right hand side term, $c \in (\frac{1}{4}, \frac{1}{2}]$ and usually achieves optimal performance with c slightly greater than $\frac{1}{4}$ in practice. This is a convex optimization problem and can be solved efficiently. The agent will choose the action i with the maximal $U_i(t)$. Algorithm 1 shows the pseudo-code of the Hellinger-UCB algorithm.

Algorithm 1 Hellinger-UCB

1. Known Parameters: T(time horizon), K(action set), $i \in K$(action), $r_t(i)$(reward given action)
2. For $t = 1$ to $|K|$:
   (a) $A_t = i = t\%|K|$
   (b) $N_i(t) = 1$
   (c) $S_i(t) = r_t(A_t)$

   end for
3. For $t = |K| + 1$ to T:
   (a) $A_t = \arg max_i \sup\{\dot{\psi}(\theta) : H^2(P_{\hat{\theta}_{i,t-1}}, P_\theta) \leqslant 1 - e^{-c\frac{\log(t)}{N_i(t)}}\}$, where $P_{\hat{\theta}_{t-1}}$ is the maximum likelihood estimation(MLE) of the reward distribution based on the past observations.
   (b) $r = r_t(A_t)$
   (c) $N_i(t) += \mathbb{I}\{A_t = i\}$
   (d) $S_i(t) += r$

   end for

3.2 Optimality of Hellinger-UCB

As a UCB-based algorithm for the stochastic multi-armed bandit problem, we are interested in whether the pseudo regret of the Hellinger-UCB algorithm is optimal. The following theorem guarantees the optimality of this algorithm. We first derive the upper bound of each sub-optimal arm's expected number of pulls.

Theorem 2 *Consider a multi-armed bandit problem with K arms and the associated payoffs are some distributions in a one-parameter exponential family. Let a^* denote the optimal arm with expectation μ^* and i denote some sub-optimal arm with expectation μ_i such that $\mu_i < \mu^*$. For any $T > 0$, the number of picks of arm i by Hellinger-UCB is $N_i(T)$. For any $\epsilon > 0$*

$$\mathbb{E}[N_i(T)] \leqslant -\frac{c\log(T)}{\log(1 - \frac{H^2(\mu^*,\mu_i)}{1+\epsilon})} + \frac{C_1(\epsilon)}{T^{C_2(\epsilon)}} + \sum_{t=1}^{T} \frac{1}{t^{2c}} + \frac{e^{-2H^2(\mu^*,\mu_i)}}{1 - e^{-2H^2(\mu^*,\mu_i)}}$$

where $C_1(\epsilon) = -\frac{c}{\log(1-\frac{H^2(\mu^*,\mu_i)}{1+\epsilon})} > 0$ *and* $C_2(\epsilon) = \frac{(\sqrt{1+\epsilon}-1)^2}{1+\epsilon} > 0$.

if $c > \frac{1}{4}$,

$$\mathbb{E}[N_i(T)] \leqslant -\frac{c\log(T)}{\log(1 - \frac{H^2(\mu^*,\mu_i)}{1+\epsilon})} + \frac{C_1(\epsilon)}{T^{C_2(\epsilon)}} + O(1)$$

Proof: See Appendix A for details of the proof □

Then the upper bound of the pseudo-regret is just a direct result of Theorem 3.1

Theorem 3 *The regret of Hellinger-UCB satisfies:*

$$\bar{R}_T \leqslant \sum_{i:\mu_i \leqslant \mu^*} \Delta_i \mathbb{E}[N_i(T)] \tag{7}$$

.

4 Numerical result

The long-run online experiment is conducted in the front page content recommendation business of JD Finance App. The recommendation system is designed to provide personalized multi-type content recommendations to the users. For each request, the cold start model is required to rank a set of articles and tweets, and then present the top-rank contents to the users. All three algorithms, UCB1, KL-UCB, and Hellinger-UCB, rank about ten thousand of contents(Financial articles) from the cold start pool with estimated CTR. CTR is modeled as the mean reward of a series of Bernoulli trials, which is the exact historical clicks and impression information. Three UCB algorithms share the whole traffic and the final impression is generated by randomly selecting one of three results uniformly. The system records 1 point as a reward to the corresponding algorithm when the user has any positive interaction (click/like/comment) with the content. Under this setting, the comparison of rewards among the three algorithms will give an insight into CTR for each algorithm. We show the upper confidence bound of UCB1, KL-UCB, and Hellinger-UCB in the appendix.

Figure 1 shows the cumulative reward plot of three algorithms in a two-month experiment from Oct. 2020 to Nov. 2020. It is very clear that Hellinger-UCB (blue line) significantly outperforms KL-UCB (orange line) and UCB1 algorithm (green line). In fact, the Hellinger-UCB algorithm achieves about 33% more clicks than the KL-UCB algorithm and almost 100% more clicks than the UCB1 algorithm. Hellinger-UCB algorithm obtains more clicks in the early period and then achieves even more clicks as the learning continues. This is an encouraging illustration of the potential power of Hellinger-UCB in real applications.

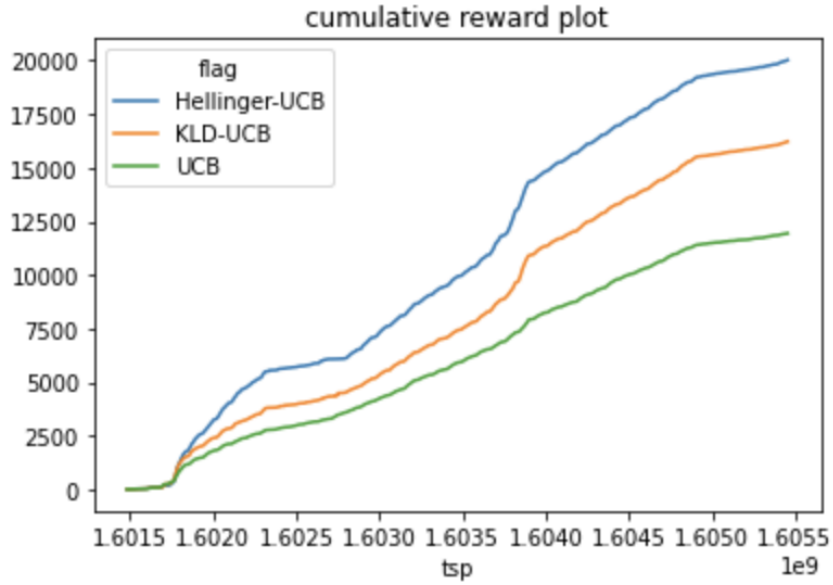

Figure 1: Cumulative reward plot of different UCB algorithms. The y-axis is reward points. The x-axis is a time stamp recorded as 9 digits integer.

5 Conclusion

We presented the Hellinger-UCB algorithm for the stochastic multi-armed bandit problem. In the case that the reward is from an unknown exponential family, we provide the detailed formula of the algorithm and an optimal regret upper bound that achieves $O(\log(T))$. We present real numerical experiments that show significant improvement over other variants of UCB algorithms. The cumulative reward form the proposed algorithm is higher. We also show the algorithm has a closed-form UCB when the reward is a bernoulli distribution, which is a beneficial property for low-latency applications.

References

[1] Rajeev Agrawal. Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in applied probability*, 27(4):1054–1078, 1995.

[2] Jean-Yves Audibert, Sébastien Bubeck, et al. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

[3] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

[4] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27, 2014.

[5] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pages 1516–1541, 2013.

[6] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

[7] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

[8] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[9] David Liau, Zhao Song, Eric Price, and Ger Yang. Stochastic multi-armed bandits in constant space. In *International Conference on Artificial Intelligence and Statistics*, pages 386–394. PMLR, 2018.

[10] Herbert E. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1952.

A Appendix

A.1 Proof of Theorem 3.1

Proof: Hellinger-UCB algorithm relies on the following upper confidence bound for μ_i:

$$u_i(t) = max\{q > \hat{\mu_i}(t) : H^2(\hat{\mu_i}(t), q) \leqslant 1 - \exp\{-c\frac{log(t)}{N_i(t)}\}\} \tag{8}$$

The expectation of $N_i(T)$ is upper-bounded by using the following decomposition. When a sub-optimal arm i is pulled, then the upper confidence bound of the optimal arm $u^*(t)$ based on historical observations is either greater or less than its true expectation μ^*. In the latter case,

$$\mathbb{E}[N_i(T)] = \mathbb{E}[\sum_{t=1}^{T} \mathbb{I}\{A_t = i\}] \tag{9}$$

$$= \mathbb{E}[\sum_{t=1}^{T} \mathbb{I}\{A_t = i, \mu^* > u^*(t)\}] + \mathbb{E}[\sum_{t=1}^{T} \mathbb{I}\{A_t = i, \mu^* \leqslant u^*(t)\}] \tag{10}$$

$$\leqslant \sum_{t=1}^{T} \mathbb{P}\{\mu^* > u^*(t)\} + \mathbb{E}[\sum_{t=1}^{T} \mathbb{I}\{A_t = i, \mu^* \leqslant u^*(t)\}] \tag{11}$$

$$\leqslant C_1(\epsilon)\log(T) + \frac{(C_2(\epsilon)H^2(\mu^*, \mu_i))^{-1}}{T^{2C_1(\epsilon)C_2(\epsilon)H^2(\mu^*,\mu_i)}} \tag{12}$$

$$+ \frac{e^{-2H^2(\mu^*,\mu_i)}}{1 - e^{-2H^2(\mu^*,\mu_i)}} + \sum_{t=1}^{T} \frac{1}{t^{2c}} \tag{13}$$

The last inequality is from Lemma 1

$$\sum_{t=1}^{T}\mathbb{P}\{\mu^* > u^*(t)\} \leqslant \sum_{t=1}^{T}\frac{1}{t^{2c}} \tag{14}$$

and Lemma 3

$$\mathbb{E}[\sum_{t=1}^{T}\mathbb{I}\{A_t = i, \mu^* \leqslant u^*(t)\}] \tag{15}$$

$$\leqslant C_1(\epsilon)\log(T) + \frac{(C_2(\epsilon)H^2(\mu^*,\mu_i))^{-1}}{T^{2C_1(\epsilon)C_2(\epsilon)H^2(\mu^*,\mu_i)}} + \frac{e^{-2H^2(\mu^*,\mu_i)}}{1-e^{-2H^2(\mu^*,\mu_i)}} \tag{16}$$

If $c > \frac{1}{4}$, according to Lemma 2 and Lemma 3, we have

$$\mathbb{E}[N_i(T)] \leqslant C_1(\epsilon)\log(T) + \frac{(C_2(\epsilon)H^2(\mu^*,\mu_i))^{-1}}{T^{2C_1(\epsilon)C_2(\epsilon)H^2(\mu^*,\mu_i)}} + O(1) \tag{17}$$

The details of these lemmas are in the following section. □

A.2 THE PROOF OF THE THEOREM

This concentration inequality[5] will be used several times

$$\mathbb{P}\{\hat{\mu}(n) > \mu, KL(\hat{\mu}(n),\mu) > \frac{f(n)}{n}\} \leqslant e^{-f(n)} \tag{18}$$

. The following lemmas support the proof of the main theorem.

Lemma 1

$$\sum_{t=1}^{T}\mathbb{P}\{\mu^* > u^*(t)\} \leqslant \sum_{t=1}^{T}\frac{1}{t^{2c}}$$

.

Proof: $\hat{\mu}^*(t)$ is the M.L.E. of μ^*, then

$$\mathbb{P}\{\mu^* > u^*(t)\} \tag{19}$$

$$\leqslant \mathbb{P}\{\mu^* > \hat{\mu}^*(t), H^2(\mu^*,\hat{\mu}^*(t)) \geqslant 1 - \exp\{-c\frac{\log(t)}{N^*(t)}\}\} \tag{20}$$

Since for exponential family $-\log(1-H^2(\mu^*,\hat{\mu}^*(t))) \leqslant \frac{1}{2}KL(\hat{\mu}^*(t),\mu^*)$, (20) becomes

$$\mathbb{P}\{\mu^* > u^*(t)\} \tag{21}$$

$$\leqslant \mathbb{P}\{\mu^* > \hat{\mu}^*(t), 1 - e^{-\frac{1}{2}KL(\hat{\mu}^*(t),\mu^*)} \geqslant (1 - \exp\{-c\frac{\log(t)}{N^*(t)}\})\} \tag{22}$$

$$\leqslant \mathbb{P}\{\mu^* > \hat{\mu}^*(t), KL(\hat{\mu}^*(t),\mu^*) \geqslant 2c\frac{\log(t)}{N^*(t)}\} \tag{23}$$

$$\leqslant e^{-2c\log(t)} \tag{24}$$

$$= \frac{1}{t^{2c}} \tag{25}$$

. Then

$$\sum_{t=1}^{T}\mathbb{P}\{\mu^* > u^*(t)\} \leqslant \sum_{t=1}^{T}\frac{1}{t^{2c}} \tag{26}$$

□

Lemma 2 *If* $c > \frac{1}{4}$ *in*

$$u^*(t) = max\{q > \hat{\mu_i}(t) : H^2(\hat{\mu^*}(t), q) \leqslant 1 - \exp\{-c\frac{log(t)}{N^*(t)}\}\}$$

then

$$\sum_{t=1}^{\infty}\mathbb{P}\{\mu^* > u^*(t)\} = O(1)$$

.

Proof: $\hat{\mu}^*(t)$ is the M.L.E. of μ^* at t

$$\mathbb{P}\{\mu^* > u^*(t)\} \tag{27}$$

$$\leqslant \mathbb{P}\{\mu^* > \hat{\mu}^*(t), H^2(\mu^*, \hat{\mu}^*(t)) \geqslant 1 - \exp\{-c\frac{\log(t)}{N^*(t)}\}\} \tag{28}$$

Since $H^2(\mu^*, \hat{\mu}^*(t)) = \frac{1}{4}\text{KL}(\hat{\mu}^*(t), \mu^*)$ if $t \to \infty$. There exist $T_1 > 0$ and $\delta_1 > 0$ such that for $t > T_1$

$$H^2(\mu^*, \hat{\mu}^*(t)) \leqslant \frac{1}{4}\text{KL}(\hat{\mu}^*(t), \mu^*) + \delta_1 \tag{29}$$

It is known that $\delta_1 = O(N^*(t)^{-\frac{3}{2}})$. Thus (28) becomes

$$\mathbb{P}\{\mu^* > u^*(t)\} \tag{30}$$

$$\leqslant \mathbb{P}\{\mu^* > \hat{\mu}^*(t), \text{KL}(\hat{\mu}^*(t), \mu^*) \geqslant 4(1 - \exp\{-c\frac{\log(t)}{N^*(t)}\} - \delta_1)\} \tag{31}$$

We have the Taylor expansion

$$\exp\{-c\frac{\log(t)}{N^*(t)}\} = 1 - c\frac{\log(t)}{N^*(t)} + \frac{1}{2!}(-c\frac{\log(t)}{N^*(t)})^2 + \frac{1}{3!}(-c\frac{\log(t)}{N^*(t)})^3 + ... \tag{32}$$

$$= 1 - c\frac{\log(t)}{N^*(t)} - \frac{\log(t)}{N^*(t)}R(c,t) \tag{33}$$

where $R(c,t) = \sum_{k=2}^{\infty} \frac{1}{k!}c^k(-\frac{\log(t)}{N^*(t)})^{k-1}$ is a negative function for $c < 1$. Notice $\lim_{t\to\infty} R(c,t) \to 0$ since $\lim_{t\to\infty} \frac{\log(t)}{N^*(t)} = 0$. There exist $T_2 > 0$ and $\delta > 0$ such that $-\delta < R(c,t) < 0$. Therefore for $t > \max(T_1, T_2)$, (31) will be

$$\mathbb{P}\{\mu^* > u^*(t)\} \tag{34}$$

$$\leqslant \mathbb{P}\{\mu^* > \hat{\mu}^*(t), \text{KL}(\hat{\mu}^*(t), \mu^*) \geqslant 4c\frac{\log(t)}{N^*(t)} + \frac{4\log(t)}{N^*(t)}R(c,t) - 4\delta_1\} \tag{35}$$

$$\leqslant \mathbb{P}\{\mu^* > \hat{\mu}^*(t), \text{KL}(\hat{\mu}^*(t), \mu^*) \geqslant 4c\frac{\log(t)}{N^*(t)} - \frac{4\log(t)}{N^*(t)}\delta - 4\delta_1\} \tag{36}$$

$$\leqslant \exp\{-4c\log(t) + 4\delta\log(t) + 4N^*(t)\delta_1\} \tag{37}$$

We now have

$$\mathbb{P}\{\mu^* > u^*(t)\} \leqslant \frac{e^{4N^*(t)\delta_1}}{t^{4(c-\delta)}} \leqslant \frac{e^{O(N^*(t)^{-\frac{1}{2}})}}{t^{4(c-\delta)}} \leqslant \frac{m_1}{t^{4(c-\delta)}} \tag{38}$$

for some finite $m_1 > 0$.

For $t \leqslant \max(T_1, T_2)$, there must exist some $m_2 > 0$ such that

$$\mathbb{P}\{\mu^* > u^*(t)\} \leqslant \frac{m_2}{t^{4(c-\delta)}} \tag{39}$$

Therefore for $M = \max(m_1, m_2)$, the following result holds

$$\mathbb{P}\{\mu^* > u^*(t)\} \leqslant \frac{M}{t^{4(c-\delta)}} \tag{40}$$

and $c > \frac{1}{4} + \delta$ implies the summation

$$\sum_{t=1}^{\infty} \mathbb{P}\{\mu^* > u^*(t)\} \leqslant \sum_{t=1}^{\infty} \frac{M}{t^{4(c-\delta)}} = O(1) \tag{41}$$

□

Lemma 3 *For any $\epsilon > 0$, then*

$$\mathbb{E}[\sum_{t=1}^{T} \mathbb{I}\{A_t = i, \mu^* \leqslant u^*(t)\}]$$

$$\leqslant C_1(\epsilon)\log(T) + \frac{(C_2(\epsilon)H^2(\mu^*, \mu_i))^{-1}}{T^{2C_1(\epsilon)C_2(\epsilon)H^2(\mu^*,\mu_i)}} + \frac{e^{-2H^2(\mu^*,\mu_i)}}{1 - e^{-2H^2(\mu^*,\mu_i)}}$$

where $C_1(\epsilon) = -\frac{c}{\log(1-\frac{H^2(\mu^*,\mu_i)}{1+\epsilon})} > 0$ and $C_2(\epsilon) = \frac{(\sqrt{1+\epsilon}-1)^2}{1+\epsilon} > 0$.

Proof: Arm i is sub-optimal with expected reward μ_i and $\hat{\mu}_i(t)$ is the M.L.E. for μ_i at t,. Then we have

$$\mathbb{E}[\sum_{t=1}^{T}\mathbb{I}\{A_t=i,\mu^*\leqslant u^*(t)\}] \tag{42}$$

$$=\mathbb{E}[\sum_{t=1}^{T}\mathbb{I}\{A_t=i,\mu^*\leqslant u^*(t),\mu^*\leqslant \hat{\mu}_i(t)\}]+ \tag{43}$$

$$\mathbb{E}[\sum_{t=1}^{T}\mathbb{I}\{A_t=i,\mu^*\leqslant u^*(t),\mu^*> \hat{\mu}_i(t)\}] \tag{44}$$

$$\leqslant C_1(\epsilon)\log(T)+\frac{(C_2(\epsilon)H^2(\mu^*,\mu_i))^{-1}}{T^{2C_1(\epsilon)C_2(\epsilon)H^2(\mu^*,\mu_i)}}+\frac{e^{-2H^2(\mu^*,\mu_i)}}{1-e^{-2H^2(\mu^*,\mu_i)}} \tag{45}$$

(45) is according to Lemma B.4 and Lemma B.5.

□

Lemma 4

$$\mathbb{E}[\sum_{t=1}^{T}\mathbb{I}\{A_t=i,\mu^*\leqslant u^*(t),\mu^*\leqslant \hat{\mu}_i(t)\}]\leqslant\frac{e^{-2H^2(\mu^*,\mu_i)}(1-e^{-2TH^2(\mu^*,\mu_i)})}{1-e^{-2H^2(\mu^*,\mu_i)}}$$

Proof:

$$\mathbb{E}[\sum_{t=1}^{T}\mathbb{I}\{A_t=i,\mu^*\leqslant u^*(t),\mu^*\leqslant \hat{\mu}_i(t)\}] \tag{46}$$

$$\leqslant\mathbb{E}[\sum_{t=1}^{T}\mathbb{I}\{A_t=i,\mu^*\leqslant \hat{\mu}_i(t),H^2(\mu^*,\mu_i)\leqslant H_2(\hat{\mu}_i(t),\mu_i)\}] \tag{47}$$

$$\leqslant\mathbb{E}[\sum_{t=1}^{T}\mathbb{I}\{A_t=i,\mu^*\leqslant \hat{\mu}_i(t),H^2(\mu^*,\mu_i)\leqslant \frac{1}{2}KL(\hat{\mu}_i(t),\mu_i)\}] \tag{48}$$

$$=\mathbb{E}[\sum_{t=1}^{T}\sum_{s=1}^{t}\mathbb{I}\{A_t=i,N_i(t)=s,\mu^*\leqslant \hat{\mu}_i(s),H^2(\mu^*,\mu_i)\leqslant \frac{1}{2}KL(\hat{\mu}_i(s),\mu_i)\}] \tag{49}$$

$$=\mathbb{E}[\sum_{s=1}^{T}\sum_{t=s}^{T}\mathbb{I}\{A_t=i,N_i(t)=s\}\mathbb{I}\{\mu^*\leqslant \hat{\mu}_i(s),H^2(\mu^*,\mu_i)\leqslant \frac{1}{2}KL(\hat{\mu}_i(s),\mu_i)\}] \tag{50}$$

$$=\mathbb{E}[\sum_{s=1}^{T}\mathbb{I}\{\mu^*\leqslant \hat{\mu}_i(s),H^2(\mu^*,\mu_i)\leqslant \frac{1}{2}KL(\hat{\mu}_i(s),\mu_i)\}\sum_{t=s}^{T}\mathbb{I}\{A_t=i,N_i(t)=s\}] \tag{51}$$

Notice in (51) $\sum_{t=s}^{T}\mathbb{I}\{A_t=i,N_i(t)=s\}]\leqslant 1$ and thus

$$\mathbb{E}[\sum_{t=1}^{T}\mathbb{I}\{A_t=i,\mu^*\leqslant u^*(t),\mu^*\leqslant \hat{\mu}_i(t)\}] \tag{52}$$

$$\leqslant\mathbb{E}[\sum_{s=1}^{T}\mathbb{I}\{\mu^*\leqslant \hat{\mu}_i(s),H^2(\mu^*,\mu_i)\leqslant \frac{1}{2}KL(\hat{\mu}_i(s),\mu_i)\}] \tag{53}$$

$$=\sum_{s=1}^{T}\mathbb{P}\{\mu^*\leqslant \hat{\mu}_i(s),H^2(\mu^*,\mu_i)\leqslant \frac{1}{2}KL(\hat{\mu}_i(s),\mu_i)\} \tag{54}$$

$$\leqslant\sum_{s=1}^{T}e^{-2sH^2(\mu^*,\mu_i)} \tag{55}$$

$$=\frac{e^{-2H^2(\mu^*,\mu_i)}(1-e^{-2TH^2(\mu^*,\mu_i)})}{1-e^{-2H^2(\mu^*,\mu_i)}} \tag{56}$$

It is easy to show

$$\lim_{T\to\infty}\mathbb{E}[\sum_{t=1}^{T}\mathbb{I}\{A_t=i,\mu^*\leqslant u^*(t),\mu^*\leqslant\hat{\mu}_i(t)\}]=\frac{e^{-2H^2(\mu^*,\mu_i)}}{1-e^{-2H^2(\mu^*,\mu_i)}} \tag{57}$$

which is a problem-dependent constant. □

Lemma 5 *For any $\epsilon>0$, then*

$$\begin{aligned}&\mathbb{E}[\sum_{t=1}^{T}\mathbb{I}\{A_t=i,\mu^*\leqslant u^*(t),\mu^*>\hat{\mu}_i(t)\}]\\ \leqslant& C_1(\epsilon)\log(T)+\frac{(C_2(\epsilon)H^2(\mu^*,\mu_i))^{-1}}{T^{2C_1(\epsilon)C_2(\epsilon)H^2(\mu^*,\mu_i)}}\end{aligned}$$

where $C_1(\epsilon)=-\frac{c}{\log(1-\frac{H^2(\mu^*,\mu_i)}{1+\epsilon})}>0$ and $C_2(\epsilon)=\frac{(\sqrt{1+\epsilon}-1)^2}{1+\epsilon}>0$.

Proof: Similar to the proof for Lemma B.4 we can have

$$\mathbb{E}[\sum_{t=1}^{T}\mathbb{I}\{A_t=i,\mu^*\leqslant u^*(t),\mu^*>\hat{\mu}_i(t)\}] \tag{58}$$

$$\leqslant\mathbb{E}[\sum_{t=1}^{T}\sum_{s=1}^{t}\mathbb{I}\{A_t=i,N_i(t)=s,H^2(\mu^*,\hat{\mu}_i(s))<1-e^{-c\frac{\log(t)}{s}}\}] \tag{59}$$

$$\leqslant\mathbb{E}[\sum_{s=1}^{T}\sum_{t=s}^{T}\mathbb{I}\{A_t=i,N_i(t)=s\}\mathbb{I}\{H^2(\mu^*,\hat{\mu}_i(s))<1-e^{-c\frac{\log(T)}{s}}\}] \tag{60}$$

$$\leqslant\mathbb{E}[\sum_{s=1}^{T}\mathbb{I}\{H^2(\mu^*,\hat{\mu}_i(s))<1-e^{-c\frac{\log(T)}{s}}\}] \tag{61}$$

$$=\sum_{s=1}^{T}\mathbb{P}\{H^2(\mu^*,\hat{\mu}_i(s))<1-e^{-c\frac{\log(T)}{s}}\} \tag{62}$$

$$\leqslant K_T+\sum_{s=K_T+1}^{T}\mathbb{P}\{H^2(\mu^*,\hat{\mu}_i(s))<1-e^{-c\frac{\log(T)}{K_T}}\} \tag{63}$$

where $K_T=C_1(\epsilon)\log(T)$ and $C_1(\epsilon)=-\frac{c}{\log(1-\frac{H^2(\mu^*,\mu_i)}{1+\epsilon})}>0$. Then substitute this into (63). Then

$$\mathbb{E}[\sum_{t=1}^{T}\mathbb{I}\{A_t=i,\mu^*\leqslant u^*(t),\mu^*>\hat{\mu}_i(t)\}] \tag{64}$$

$$\leqslant C_1(\epsilon)\log(T)+\sum_{s=K_T+1}^{T}\mathbb{P}\{H^2(\mu^*,\hat{\mu}_i(s))<\frac{H^2(\mu^*,\mu_i)}{1+\epsilon}\} \tag{65}$$

There exist $\mu'\in(\mu_i,\mu^*)$ such that $(1+\epsilon)H^2(\mu^*,\mu')=H^2(\mu^*,\mu_i)$. Then $H^2(\mu^*,\hat{\mu}_i(s))<\frac{H^2(\mu^*,\mu_i)}{1+\epsilon}$ implies $\mu_i<\mu'<\hat{\mu}_i(s)$ and $H^2(\mu_i,\hat{\mu}_i(s))>H^2(\mu_i,\mu')$. The second term in (65)

becomes

$$\sum_{s=K_T+1}^{\infty} \mathbb{P}\{H^2(\mu^*, \hat{\mu}_i(s)) < \frac{H^2(\mu^*, \mu_i)}{1+\epsilon}\} \tag{66}$$

$$\leqslant \sum_{s=K_T+1}^{\infty} \mathbb{P}\{\hat{\mu}_i(s) > \mu_i, H^2(\mu_i, \hat{\mu}_i(s)) > H^2(\mu_i, \mu')\} \tag{67}$$

$$\leqslant \sum_{s=K_T+1}^{\infty} \mathbb{P}\{\hat{\mu}_i(s) > \mu_i, \frac{1}{2}KL(\hat{\mu}_i(s), \mu_i) > H^2(\mu_i, \mu')\} \tag{68}$$

$$\leqslant \sum_{s=K_T+1}^{\infty} e^{-2sH^2(\mu_i,\mu')} \tag{69}$$

$$= \frac{e^{-2(K_T+1)H^2(\mu_i,\mu')}}{1-e^{-2H^2(\mu_i,\mu')}} \tag{70}$$

The numerator of (70) $e^{-2(K_T+1)H^2(\mu_i,\mu')} \leqslant T^{-2C_1(\epsilon)H^2(\mu_i,\mu_\prime)}$. For the denominator of (70), we have $1-e^{-2H^2(\mu_i,\mu')} > H^2(\mu_i,\mu')$ since $1-e^{-x} = 1-(1-x+\frac{x^2}{2}-...) > x-\frac{x^2}{2} > x-\frac{x}{2}$ if $0<x<1$. Therefore

$$\sum_{s=K_T+1}^{\infty} \mathbb{P}\{H^2(\mu^*, \hat{\mu}_i(s)) < \frac{H^2(\mu^*, \mu_i)}{1+\epsilon}\} \leqslant \frac{H^2(\mu_i,\mu')^{-1}}{T^{2C_1(\epsilon)H^2(\mu_i,\mu_\prime)}} \tag{71}$$

Since the squared Hellinger distance is a metric, we have

$$H(\mu^*, \mu_i) = \sqrt{1+\epsilon} H((\mu^*, \mu')) \tag{72}$$

$$\geqslant \sqrt{1+\epsilon}(H(\mu^*, \mu_i) - H(\mu', \mu_i)) \tag{73}$$

This implies

$$H^2(\mu', \mu_i) \geqslant \frac{(\sqrt{1+\epsilon}-1)^2}{1+\epsilon} H^2(\mu^*, \mu_i) = C_2(\epsilon) H^2(\mu^*, \mu_i) \tag{74}$$

Finally, we conclude that

$$\mathbb{E}[\sum_{t=1}^{T} \mathbb{I}\{A_t = i, \mu^* \leqslant u^*(t), \mu^* > \hat{\mu}_i(t)\}] \leqslant \frac{(C_2(\epsilon)H^2(\mu^*, \mu_i))^{-1}}{T^{2C_1(\epsilon)C_2(\epsilon)H^2(\mu^*,\mu_i)}} \tag{75}$$

where $C_1(\epsilon) > 0$ and $C_2(\epsilon) > 0$. □

B Appendix

B.1 Solution of different UCBs

We will compare the solutions of three UCB algorithms and introduce the advantages and disadvantages of those algorithms. We will also prove that Hellinger-UCB has a close form with binomial distribution assumption of rewards.

B.1.1 UCB1 algorithm

The UCB1 algorithm[3], regardless of the reward distribution, always uses the following UCB formula:

$$U_t(i) = \hat{\mu}_{i,t-1} + \sqrt{\frac{2\log(t)}{N_i(t)}} \tag{76}$$

where $\hat{\mu}_{i,t-1}$ is the success ratio of content i at time $t-1$, $N_i(t)$ the number of impressions of content i at time t. UCB1 algorithm can always compute its confidence bound with a counting process.

B.1.2 KL-UCB algorithm

The KL-UCB algorithm solves the following optimization problem numerically to find the best arm

$$U_t(i) = \sup\{\mu(\theta) : KL(P_{\hat{\theta}_{t-1}}|P_\theta) \leqslant \frac{\log(t) + c\log\log(t)}{N_i(t)}\} \tag{77}$$

where $P_{\hat{\theta}_{t-1}}$ and $P_{\hat{\theta}_t}$ are the reward distribution at $t-1$ and t. With the assumption of reward distribution in **??**, the solution of KL-UCB becomes the following root-finding problem

$$U_t(i) = \sup\{p_t : p_{t-1}\log\frac{p_{t-1}}{p_t} + (1-p_{t-1})\log\frac{1-p_{t-1}}{1-p_t} = C\} \tag{78}$$

where

$$C = \frac{\log(t) + c\log\log(t)}{N_i(t)} \tag{79}$$

It is easy to get that 78 does not have a closed-form solution and requires a numerical solver to iterate the solution. Therefore KL-USB requires much more computation resources than UCB1 though KL-UCB has better performance than UCB1. In real-world applications, KL-UCB is less favorable than UCB1 since it requires much more careful engineering controls.

B.1.3 Hellinger-UCB algorithm

The Hellinger-UCB chooses the best arm by solving the following optimization problem

$$U_t(i) = \sup\{\dot{\psi}(\theta) : H^2(P_{\hat{\theta}_{i,t-1}}, P_\theta) \leqslant 1 - e^{-c\frac{\log(t)}{N_i(t)}}\} \tag{80}$$

unlike KL-UCB, the Hellinger-UCB algorithm has a closed-form solution with the binomial reward.

Recall that the squared Hellinger distance between two Binomial distributions $B(n,p)$ and $B(n,q)$ is given by

$$H^2(p,q) = 1 - \sqrt{(1-p)(1-q)} - \sqrt{pq} \tag{81}$$

Let $f(q) = 1 - \sqrt{(1-p)(1-q)} - \sqrt{pq}$, its derivative is $f'(q) = \sqrt{\frac{1-p}{1-q}} - \sqrt{\frac{p}{q}}$. It is easy to see that

$$f'(q) : \begin{cases} < 0, & q < p, \\ = 0, & q = p, \\ > 0, & q > p. \end{cases} \tag{82}$$

The solution to the above equation 80 must be on the squared Hellinger ball. Let R be the radius of the squared Hellinger Ball, i.e.

$$R = 1 - \sqrt{(1-p)(1-q)} - \sqrt{pq}. \tag{83}$$

Divide both sides by $\sqrt{q}$ and let $m_1 = \sqrt{\frac{1-p}{p}}$ and $m_2 = \frac{1-R}{\sqrt{p}}$

$$\sqrt{q} + m_1\sqrt{1-q} = m_2. \tag{84}$$

Take the square of both sides and simplify the equation

$$2m_1\sqrt{q(1-q)} = m_2^2 - m_1^2 + (m_1^2 - 1)q \tag{85}$$

Repeat above procedure one more time and simplify the result

$$(m_1^2+1)^2q^2 + 2(m_1^2m_2^2 - m_1^4 - m_1^2 - m_2^2)q + (m_2^2 - m_1^2)^2 = 0. \tag{86}$$

Let $a = (m_1^2+1)^2$, $b = 2(m_1^2m_2^2 - m_1^4 - m_1^2 - m_2^2)$ and $c = (m_2^2 - m_1^2)^2$, the root of 86 is

$$q = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{87}$$

The larger root is desired. Therefore, Hellinger-UCB has a close form solution with binomial reward distribution assumption.