A Note on KL-UCB+ Policy for the Stochastic Bandit

Junya Honda*
The University of Tokyo / RIKEN

Abstract

A classic setting of the stochastic \(K\)-armed bandit problem is considered in this note. In this problem it has been known that KL-UCB policy achieves the asymptotically optimal regret bound and KL-UCB+ policy empirically performs better than the KL-UCB policy although the regret bound for the original form of the KL-UCB+ policy has been unknown. This note demonstrates that a simple proof of the asymptotic optimality of the KL-UCB+ policy can be given by the same technique as those used for analyses of other known policies.

1 Overview

In the problem of the stochastic bandit problems, it is known that there exists a (problem-dependent) regret lower bound \([1,2]\). It can be achieved by, for example, the DMED policy \([3]\) for the model of nonparametric distributions over \([0,1]\). On the other hand, the policy proposed in \([2]\), which is often called KL-UCB \([5]\), achieves the asymptotic bound for some models such as one-parameter exponential families and the family of distributions over a finite support. One of the conference version \([6]\) of \([5]\) also proposed KL-UCB+ policy, which empirically performs better than KL-UCB but does not have a theoretical guarantee.

The KL-UCB+ policy is obtained by replacing \(t\) in the confidence bound with \(t/N_i(t)\), where \(t\) is the current number of trials and \(N_i(t)\) is the number of samples from the arm \(i\). It is discussed in \([6]\] that KL-UCB+ can be related to DMED, which also has term \(t/N_i(t)\) in the criterion for choosing the arm to pull, and \([6]\] used the name DMED+ for the same policy based on this observation.

After these works, IMED policy is proposed \([7]\] as an index-policy version of the DMED in \([3]\) (or equivalently, DMED+ in \([5]\)), which achieves the asymptotic bound and empirically performs almost the same as the KL-UCB+ policy with low computational cost. On the other hand, the asymptotic optimality of the KL-UCB+ has not been proved to the best of the author’s knowledge, although proofs are given for its variants \([8,9]\).

In this note, we show that this difference, known optimality of DMED/IMED and unknown optimality of KL-UCB+, simply comes from the difference of the used techniques between \([3,7]\) and \([5]\) by demonstrating that the asymptotic optimality of (a slightly generalized version of) KL-UCB+ can be proved by exactly the same argument as those for DMED and IMED. To be more specific, whereas the typical technique for KL-UCB such as \([5,10]\) reduces the regret analysis to the evaluation of a hitting probability of a boundary for some stochastic process, the technique in \([3,7]\) reduces it to the evaluation of the expected waiting-time for the optimal arm to be pulled the next time. These two analyses incur different looseness arising from double-counting of events and there is no clear winner, but the latter one seems to be convenient for the proof of the asymptotic optimality of KL-UCB+.

Note that the original analyses in \([3]\) and \([7]\) are given for general distributions over \([0,1]\) and \((−\infty,1]\), respectively, but this note only considers Bernoulli distributions to simply illustrate the difference of the regret decomposition between \([3,7]\) and \([5]\).

*Email: honda@stat.t.u-tokyo.ac.jp

1The original paper by Burnetas and Katehakis \([2]\) did not explicitly give a name for the policy but it is referred to as Inflated Sample Mean (ISM) policy by their group \([4]\).
Algorithm 1: KL-UCB(α) Policy

Parameter: α ≥ 0.
1 Pull each arm once.
2 for t = K + 1, K + 2, · · · , T do
3 Pull arm i(t) = argmaxi∈[K] ̂ µi(t).

It should be noted that this result does not mean that for any model KL-UCB+ always achieves the asymptotic bound only if KL-UCB does. In fact, it seems to be almost hopeless to expect the asymptotic optimality of KL-UCB+ for general multi-parameter models with a non-compact parameter space from the discussions in [3] and [10].

2 Problem Setting

Let there be K ∈ N arms. The agent pulls one arm i(t) ∈ {1, 2, · · · , K} and observe reward r(t) ∈ {0, 1} at each round t = 1, 2, . . . . The observed reward from the arm i independently follows Bernoulli distribution Ber(µi), where the expected reward µi ∈ [0, 1] is unknown to the agent.

The objective of the agent is to maximize the cumulative reward through T ∈ N rounds, and its performance is measured by regret or pseudo-regret, which is given by

\[ \text{regret}(T) = \sum_{t=1}^{T} (\mu^* - \mu_{i(t)}), \]

where \( \mu^* = \mu_{i^*} \) for \( i^* = \text{argmax}_{i \in \{1, 2, \ldots, K\}} \mu_i \). It is shown in [1] that any policy satisfying a mild regularity condition called consistency has regret lower bound

\[ \liminf_{T \to \infty} \frac{\mathbb{E} [\text{regret}(T)]}{\log T} \geq \sum_{i \neq i^*} \frac{\mu^* - \mu_i}{d(\mu_i, \mu^*)}, \]

where \( d(x, y) = x \log(x/y) + (1-x) \log((1-x)/(1-y)) \) is the KL divergence between Bernoulli distributions Ber(x) and Ber(y).

Let \( N_i(t) \) denote the number of rounds that the arm i is pulled before the t-th round. We define \( \hat{\mu}_i(t) = \frac{1}{N_i(t)} \sum_{t'=1}^{t-1} \mathbb{I}[i(t') = i] r(t') \) as the empirical mean of the rewards from arm i before the t-th round and \( \hat{\mu}_{i,n} \) as the empirical mean of the rewards from the arm i when the arm i is pulled n times. Then we have \( \hat{\mu}_{i,N_i(t)} = \hat{\mu}_i(t) \).

We consider a slightly generalized version of KL-UCB+ policy with parameter \( \alpha \geq 0 \), which we denote by KL-UCB(\( \alpha \)) policy. In this policy, the pulled arm is the one maximizing the UCB score given by

\[ \overline{\mu}_i(t) = \sup \left\{ \mu \in [0, 1] : d(\hat{\mu}_i(t), \mu) \leq \frac{\log(T)^2}{N_i(t)} \right\}. \]

More formally, the KL-UCB(\( \alpha \)) policy is given by Algorithm 1. Here \( \alpha = 0 \) corresponds to the KL-UCB policy and \( \alpha = 1 \) corresponds to the KL-UCB+ policy. Here note that [8] proved the asymptotic optimality for the modified version with an extra exploration term such that \( t/((N_i(t))^\alpha) \) is replaced with \( (t \log^c t)/N_i(t) \) with \( c \geq 7 \).

3 Result

In this note we prove the following regret bound.
**Theorem 1.** Assume $\mu^* < 1$. The expected regret of the KL-UCB($\alpha$) policy is bounded for any $\epsilon \in (0, (\mu^* - \max_{i \neq i^*} \mu_i)/2)$ by

$$
E[\text{regret}(T)] \leq \sum_{i \neq i^*} \left( \mu^* - \mu_i \right) \left( n_i + \frac{1}{2\epsilon^2} \right) + \epsilon \Gamma(\alpha + 2) \left( 1 + \log \frac{1}{1 - \mu^*} \right) \left( \frac{\mu^*(1 - \mu^* + \epsilon)}{\epsilon^2} \right)^{\alpha+2},
$$

(1)

where $\Gamma(\cdot)$ is the gamma function and $n_i = \sup\{x \geq 0 : x^\alpha e^{x d(\mu_i + \epsilon, \mu^* - \epsilon)} \leq T\}$.

**Remark 1.** The proof only uses the fact that the reward is in $[0, 1]$ as a property of Bernoulli distributions. Thus the same regret bound holds for general distributions on $[0, 1]$ like the KL-UCB policy.

**Remark 2.** The first term $n_i$ in (1) is trivially bounded by

$$n_i \leq \frac{\log T}{d(\mu_i + \epsilon, \mu^* - \epsilon)},$$

which leads to the asymptotic bound

$$\limsup_{T \to \infty} \frac{E[\text{regret}(T)]}{\log T} \leq \sum_{i \neq i^*} \frac{\mu^* - \mu_i}{d(\mu_i, \mu^*)}.$$

We can also express $n_i$ as

$$n_i = \frac{\alpha}{d(\mu_i + \epsilon, \mu^* - \epsilon)} W_0 \left( \frac{T^\frac{\alpha}{d(\mu_i + \epsilon, \mu^* - \epsilon)}}{d(\mu_i + \epsilon, \mu^* - \epsilon)} \right),$$

where $W_0(x)$ is Lambert W function, that is, the solution of $x = ze^z$, $z \geq -1$. The expansion $W_0(x) = \log x - \log \log x + O\left( \frac{\log \log x}{\log x} \right)$ [11, Sect. 4.13] for large $x$ leads to the bound

$$n_i = \frac{\log T - \alpha \log \log T}{d(\mu_i + \epsilon, \mu^* - \epsilon)} + O \left( \frac{\log \log T}{\log T} \right).$$

(2)

Note that the derivation of (1) is essentially the same as [7, Theorem 3] for the IMED policy, which suffers from heavy dependence on $\epsilon$. On the other hand, [7, Theorem 5] gives a more complicated but tighter bound for this policy. We can expect that the same technique combined with (2) can be used to obtain a higher-order bound

$$\limsup_{T \to \infty} \frac{E[\text{regret}(T)] - \sum_{i \neq i^*} \frac{(\mu^* - \mu_i) \log T}{d(\mu_i, \mu^*)}}{\log \log T} < 0$$

for the KL-UCB($\alpha$) policy with $\alpha > 0$, although it makes the analysis much longer.

**Proof.** First we bound the regret as

$$\text{regret}(T) = \sum_{i=1}^{T} \left( \mu^* - \mu_{i(t)} \right) \leq \sum_{i \neq i^*} \left( \mu^* - \mu_i \right) \sum_{t=1}^{T} \mathbb{1}[i(t) = i, \mu^*(t) \geq \mu^* - \epsilon] + \sum_{t=1}^{T} \mathbb{1}[\mu^*(t) < \mu^* - \epsilon],$$

(3)

where we write $\mu^*(t) = \max_{i \in \{1, 2, \ldots, K\}} \mu_i(t)$. 

3
The term (A) is expressed as

$$(A) = \sum_{n=1}^{T} \sum_{t=1}^{T} \mathbb{I}[i(t) = i, \overline{\mu}^*(t) \geq \mu^* - \epsilon, N_i(t) = n]$$

where we used the fact that the event $\{N_i(t) = n, i(t) = i\}$ occurs at most once for each $(n, i)$. Since $i(t) = i$ implies $\overline{\mu}_i(t) = \overline{\mu}^*(t)$, we have

$$(A) \leq \sum_{n=1}^{T} \sum_{t=1}^{T} \mathbb{I}\left[ \bigcup_{t=1}^{T} \{i(t) = i, \overline{\mu}^*(t) \geq \mu^* - \epsilon, N_i(t) = n\} \right]$$

$$(A) \leq \sum_{n=1}^{T} \left( \sum_{t=1}^{T} \left( \sup_{\mu} \left\{ \mu : d(\hat{\mu}_i(t), \mu) \leq \frac{\log \frac{t}{N_i(t)}}{N_i(t)} \right\} \geq \mu^* - \epsilon, N_i(t) = n \right) \right)$$

$$(A) \leq \sum_{n=1}^{T} \left( \sum_{t=1}^{T} \left( \sup_{\mu} \left\{ \mu : d(\hat{\mu}_i, \mu) \leq \frac{\log \frac{t}{N_i(t)}}{n} \right\} \right) \geq \mu^* - \epsilon \right)$$

$$(A) \leq n_i + \sum_{n=1}^{T} \left( \sum_{t=1}^{T} \mathbb{I}\left[ \{d(\hat{\mu}_i, \mu^* - \epsilon) \leq d(\mu_i + \epsilon, \mu^* - \epsilon)\} \cup \{\hat{\mu}_i \geq \mu^* - \epsilon\} \right] \right)$$

(by definition $n_i = \sup \{x \geq 0 : x^\alpha e^{-\alpha x} \leq T\}$)

$$\leq n_i + \sum_{n=1}^{T} \left( \sum_{t=1}^{T} \mathbb{I}\left[ \hat{\mu}_i \geq \mu_i + \epsilon \right] \right)$$

Therefore, by Hoeffding’s inequality we have

$$\mathbb{E}(A) \leq n_i + \sum_{n=1}^{\infty} \mathbb{Pr}[\hat{\mu}_i \geq \mu_i + \epsilon]$$

$$\leq n_i + \sum_{n=1}^{\infty} e^{-2n\epsilon^2}$$

$$= n_i + \frac{1}{e^{2\epsilon^2} - 1} \leq n_i + \frac{1}{2e^2},$$

where we used the fact $e^x \geq 1 + x$.

Next we evaluate the term (B). By the definition of the UCB score we have

$$\{\overline{\mu}_i^*(t) < \mu^* - \epsilon\} \Leftrightarrow \left\{ \sup_{\mu} \left\{ \mu : d(\hat{\mu}_i^*(t), \mu) \leq \frac{\log \left( \frac{1}{N_i^*(t)} \right)}{N_i^*(t)} \right\} < \mu^* - \epsilon \right\}$$

$$= \left\{ d(\hat{\mu}_i^*(t), \mu^* - \epsilon) > \frac{\log \left( \frac{t}{N_i^*(t)} \right)}{N_i^*(t)}, \hat{\mu}_i^*(t) < \mu_i^* - \epsilon \right\}$$

$$\Leftrightarrow \left\{ t < (\log \left( \frac{t}{N_i^*(t)} \right)) \alpha e^{-\alpha (\hat{\mu}_i^*(t), \mu^* - \epsilon)}; \hat{\mu}_i^*(t) < \mu_i^* - \epsilon \right\}.$$
In this proof we used two properties specific to the Bernoulli distributions: the existence of variances, we can show by evaluating a term similar to (5) that the original form of KL-UCB policy [4]

\[ \text{the policy to have a regret bound. In fact, for the case of normal distributions with unknown means and lower bound of some policies. Therefore the boundedness of the expectation of (5) is quite essential for evaluating for the expectation of (5) such as [7]. Note that a term similar to (5) also appears in the regret for multi-parameter models with non-compact parameter spaces, where we need more sophisticated evaluation for the expectation of (5). At least one of them does not hold for most models, which make the analysis difficult especially for multi-parameter models with non-compact parameter spaces, where we need more sophisticated evaluation for the expectation of (5).}

\[ \text{We complete the proof by combining (4) and (8) with (3).} \]

**Remark 3.** In this proof we used two properties specific to the Bernoulli distributions: the existence of constant \( C(\mu, \mu') > 0 \) such that \( \inf_{x \leq \mu} \{d(x, \mu') - d(x, \mu)\} \geq C(\mu, \mu') \) in [6], and finiteness of \( \sup_x d(x, \mu) \) in [7]. At least one of them does not hold for most models, which makes the analysis difficult especially for multi-parameter models with non-compact parameter spaces, where we need more sophisticated evaluation for the expectation of [5] such as [7]. Note that a term similar to [5] also appears in the regret lower bound of some policies. Therefore the boundedness of the expectation of [5] is quite essential for the policy to have a regret bound. In fact, for the case of normal distributions with unknown means and variances, we can show by evaluating a term similar to [7] that the (original form of) KL-UCB policy [4] and Thompson sampling with Jeffreys prior [12] incur polynomial regret.
Acknowledgement

The author thanks Mr. Kohei Takagi for his survey on the analysis of KL-UCB+. The author also thanks Professor Vincent Y. F. Tan for finding many typos in the first version.

References

[1] T. L. Lai and H. Robbins, “Asymptotically efficient adaptive allocation rules,” Advances in Applied Mathematics, vol. 6, pp. 4–22, 1985.

[2] A. N. Burnetas and M. N. Katehakis, “Optimal adaptive policies for sequential allocation problems,” Advances in Applied Mathematics, vol. 17, no. 2, pp. 122–142, 1996.

[3] J. Honda and A. Takemura, “An asymptotically optimal bandit algorithm for bounded support models,” in Proceedings of COLT 2010, 2010, pp. 67–79.

[4] W. Cowan, J. Honda, and M. N. Katehakis, “Normal bandits of unknown means and variances,” Journal of Machine Learning Research, vol. 18, no. 1, pp. 5638–5665, 2017.

[5] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz, “Kullback-Leibler upper confidence bounds for optimal sequential allocation,” Annals of Statistics, vol. 41, no. 3, pp. 1516–1541, 2013.

[6] A. Garivier and O. Cappé, “The KL-UCB algorithm for bounded stochastic bandits and beyond,” in Proceedings of COLT 2011, Budapest, Hungary, 2011.

[7] J. Honda and A. Takemura, “Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards,” Journal of Machine Learning Research, vol. 16, pp. 3721–3756, 2015.

[8] E. Kaufmann, “On Bayesian index policies for sequential resource allocation,” Annals of Statistics, vol. 46, no. 2, pp. 842–865, 04 2018.

[9] P. Ménard and A. Garivier, “A minimax and asymptotically optimal algorithm for stochastic bandits,” in Proceedings of ALT 2017, 2017, pp. 223–237.

[10] O.-A. Maillard, “Boundary crossing probabilities for general exponential families,” Mathematical Methods of Statistics, vol. 27, no. 1, pp. 1–31, Jan 2018.

[11] F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, NIST Handbook of Mathematical Functions, 1st ed. New York: Cambridge University Press, 2010.

[12] J. Honda and A. Takemura, “Optimality of Thompson sampling for Gaussian bandits depends on priors,” in Proceedings of AISTATS 2014, 2014, pp. 375–383.