A Note on KL-UCB

TSAI MENG YU
Department of Computer Science
National Chiao Tung University

1. Introduction

1.1. Abstract

The multi-armed bandit problem is a simple, archetypal setting of reinforcement learning, where an agent facing a slot machine with several arms tries to maximize her profit by a proper choice of arm draws.

This paper presents a finite-time analysis of the KL-UCB algorithm, an online, horizonfree index policy for stochastic bandit problems. We prove two distinct results: first, for arbitrary bounded rewards, the KL-UCB algorithm satisfies a uniformly better regret bound than UCB and its variants; second, in the special case of Bernoulli rewards, it reaches the lower bound of Lai and Robbins.

KL-UCB is also the only method that always performs better than the basic UCB policy.

1.2. First Family of Bandit Problems

In the first family, the distribution of X_t given $A_t = a$ is assumed to belong to a family $\{p_\theta, \theta \in \Theta_a\}$ of probability distributions. Lai and Robbins (1985) proved a lower-bound on the performance of any policy, and determined optimal policies. This framework was extended to multi-parameter models by Burnetas and Katehakis (1997) who showed that the number of draws up to time n, $N_a(n)$, of any sub-optimal arm a is lower-bounded by

$$N_a(n) \geq \left(\frac{1}{\inf\limits_{\theta \in \Theta_a : \mathbb{E}[p_\theta] > \mu_{a^*}} KL(p_{\theta_a}, p_\theta)} + o(1) \right) \log(n), \tag{1}$$

where KL denotes the Kullback-Leibler divergence and $\mathbb{E}[p_\theta]$ is the expectation under p_θ; hence, the regret is lower-bounded as follows:

$$\liminf_{n \to \infty} \frac{\mathbb{E}[R_n]}{\log(n)} \geq \sum_{a : \mu_a < \mu_{a^*}} \frac{\mu_{a^*} - \mu_a}{\inf\limits_{\theta \in \Theta_a : \mathbb{E}[p_\theta] > \mu_{a^*}} KL(p_{\theta_a}, p_\theta)}. \tag{2}$$

1.3. Second Family of Bandit Problems

In the second family of bandit problems, the rewards are only assumed to be bounded (say, between 0 and 1), and policies rely directly on the estimates of the expected rewards for each arm.

The KL-UCB algorithm considered in this paper is primarily meant to address this second, non-parametric, setting.

These policies not only compute an estimate of the expected rewards, but rather an upper-confidence bound (UCB), and the agent's choice is an arm with highest UCB.

Preprint. Under review.

UCB is an online, horizon-free procedure for Auer et al. (2002) proves that there exists a constant C such that

$$\mathbb{E}[R_n] \leq \sum_{a:\mu_a<\mu_{a^*}} \frac{8\log(n)}{\mu_{a^*} - \mu_a} + C$$

UCB2 variant relies on a parameter α that needs to be tuned, depending in particular on the horizon, and satisfies the tighter regret bound

$$\mathbb{E}[R_n] \leq \sum_{a:\mu_a<\mu_{a^*}} \frac{(1+\epsilon(\alpha))\log(n)}{2(\mu_{a^*} - \mu_a)} + C(\alpha)$$

In a latter work, Audibert et al. (2009) proposed a related policy, called UCB-V, which uses an empirical version of the Bernstein bound to obtain refined upper confidence bounds. Recently, Audibert and Bubeck (2010) Audibert and Bubeck (2010) introduced an improved UCB algorithm, termed MOSS, which achieves the distribution-free optimal rate.

2. Problem Formulation

2.1. Bandit Problem

The agent sequentially chooses, for t = 1, 2, . . . , n, an arm $A_t \in \{1, ..., K\}$, and receives a reward X_t such that, conditionally on the arm choices $A_1, A_2, ...$, the rewards are independent and identically distributed, with expectation $\mu_{A_1}, \mu_{A_2}, ...$ Her policy is the (possibly randomized) decision rule that, to every past observations $(A_1, X_1, ..., A_{t-1}, X_{t-1})$, associates her next choice A_t. The best choice is any arm a^* with maximal expected reward μ_{a^*} . The performance of her policy can be measured by the regret R_n defined as the difference between the rewards she accumulates up to the horizon t = n, and the rewards that she would have accumulated during the same period, had she known from the beginning which arm had the highest expected reward.

We consider the following bandit problem: the set of actions is $\{1, ..., K\}$, where K denotes a finite integer. For each $a \in \{1, ..., K\}$, the rewards $(X_a, t)_{t\geq 1}$ are independent and bounded in $\Theta = [0, 1]$ with common expectation μ_a. The sequences $(X_{a,.})_a$ are independent. At each time step $t = 1, 2, \ldots,$ the agent chooses an action A_t according to his past observations (possibly using some independent randomization) and gets the reward $X_t = X_{A_t,N_{A_t}(t)}$, where $N_a(t) = \sum_{s=1}^{t} \mathbb{I}\{A_s = a\}$ denotes the number of times action a was chosen up to time t. The sum of rewards she has obtained when choosing action a is denoted by $S_a(t) = \sum_{s\leq t} \mathbb{I}\{A_s = a\}X_s = \sum_{s=1}^{N_a(t)} X_s = \sum_{s=1}^{N_a(t)} X_{a,s}$. For $(p, q) \in \Theta^2$ denote the Bernoulli KL divergence by

$$d(p, q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

2.2. KL-UCB Algorithm

Select action s.t.

$$\max_a \left\{ q \in \Theta : N[a]d\left(\frac{S[a]}{N[a]}, q\right) \leq \log(t) + c\log(\log(t)) \right\}$$

3. Theoretical Analysis

Theorem 1 *Consider a bandit problem with K arms and independent rewards bounded in $[0, 1]$, and denote by a^* an optimal arm. Choosing $c = 3$, the regret of the KL-UCB algorithm satisfies:*

$$\limsup_{n\to\infty} \frac{\mathbb{E}[R_n]}{\log(n)} \leq \sum_{a:\mu_a<\mu_{a^*}} \frac{\mu_{a^*} - \mu_a}{d(\mu_a, \mu_{a^*})} .$$

Algorithm 1 KL-UCB

Require: n (horizon), K (number of arms), REWARD (reward function, bounded in $[0,1]$)

for $t = 1$ **to** K **do**
  $N[t] \leftarrow 1$
  $S[t] \leftarrow \text{REWARD}(\text{arm} = t)$
end for
for $t = K+1$ **to** n **do**
  $a \leftarrow \arg\max_{1 \le a \le K} \max \left\{ q \in \Theta : N[a]\, d\left(\frac{S[a]}{N[a]}, q\right) \le \log(t) + c\log(\log(t)) \right\}$
  $r \leftarrow \text{REWARD}(\text{arm} = a)$
  $N[a] \leftarrow N[a] + 1$
  $S[a] \leftarrow S[a] + r$
end for

Theorem 2 *Consider a bandit problem with K arms and independent rewards bounded in $[0,1]$. Let $\epsilon > 0$, and take $c = 3$ in Algorithm 1. Let a^* denote an arm with maximal expected reward μ_{a^*}, and let a be an arm such that $\mu_a < \mu_{a^*}$. For any positive integer n, the number of times algorithm KL-UCB chooses arm a is upper-bounded by*

$$\mathbb{E}[N_n(a)] \le \frac{\log(n)}{d(\mu_a, \mu_{a^*})}(1+\epsilon) + C_1 \log(\log(n)) + \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}} \ ,$$

where C_1 denotes a positive constant and where $C_2(\epsilon)$ and $\beta(\epsilon)$ denote positive functions of ϵ. Hence,

$$\limsup_{n\to\infty} \frac{\mathbb{E}[N_n(a)]}{\log(n)} \le \frac{1}{d(\mu_a, \mu_{a^*})} \ .$$

Proof Consider a positive integer n, a small $\epsilon > 0$, an optimal arm a^* and a sub-optimal arm a such that $\mu_a < \mu_{a^*}$. Without loss of generality, we will assume that $a^* = 1$. For any arm b, the past average performance of arm b is denoted by $\hat{\mu}_b(t) = S_b(t)/N_b(t)$; by convenience, for every positive integer s we will also denote $\hat{\mu}_{b,s} = (X_{b,1} + \cdots + X_{b,s})/s$, so that $\hat{\mu}_t(b) = \hat{\mu}_{b,N_b(t)}$. KL-UCB relies on the following upper-confidence bound for μ_b:

$$u_b(t) = \max\left\{ q > \hat{\mu}_b(t) : N_b(t)\, d\left(\hat{\mu}_b(t), q\right) \le \log(t) + 3\log(\log(t)) \right\} \ .$$

For $x, y \in [0,1]$, define $d^+(x,y) = d(x,y)\mathbb{1}_{x<y}$. The expectation of $N_n(a)$ is upper-bounded by using the following decomposition:

$$\begin{aligned}
\mathbb{E}[N_n(a)] &= \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\{A_t = a\}\right] \le \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\{\mu_1 > u_1(t)\}\right] + \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\{A_t = a, \mu_1 \le u_1(t)\}\right] \\
&\le \sum_{t=1}^{n} \mathbb{P}\left(\mu_1 > u_1(t)\right) + \mathbb{E}\left[\sum_{s=1}^{n} \mathbb{1}\{s d^+(\hat{\mu}_{a,s}, \mu_1) < \log(n) + 3\log(\log(n))\}\right] \ ,
\end{aligned}$$

where the last inequality is a consequence of Lemma 3. The first summand is upper-bounded as follows: by Theorem 5,

$$\begin{aligned}
P(\mu_1 > u_1(t)) &\le e \left\lceil \log(t)\left(\log(t) + 3\log(\log(t))\right) \right\rceil \exp(-\log(t) - 3\log(\log(t))) \\
&= \frac{e\left\lceil \log(t)^2 + 3\log(t)\log(\log(t)) \right\rceil}{t\log(t)^3} \ .
\end{aligned}$$

Hence,

$$\sum_{t=1}^{n} P(\mu_1 > u_1(t)) \le \sum_{t=1}^{n} \frac{e\left\lceil \log(t)^2 + 3\log(t)\log(\log(t)) \right\rceil}{t\log(t)^3} \le C_1' \log(\log(n))$$

for some positive constant C_1' ($C_1' \le 7$ is sufficient). For the second summand, let

$$K_n = \left\lfloor \frac{1+\epsilon}{d^+(\mu_a, \mu_1)} \Big(\log(n) + 3\log(\log(n)) \Big) \right\rfloor \ .$$

Then:

$$\sum_{s=1}^{n} \mathbb{P}\big(sd^+(\hat{\mu}_{a,s},\mu_1) < \log(n) + 3\log(\log(n))\big)$$

$$\leq K_n + \sum_{s=K_n+1}^{\infty} \mathbb{P}\Big(sd^+(\hat{\mu}_{a,s},\mu_1) < \log(n) + 3\log(\log(n))\Big)$$

$$\leq K_n + \sum_{s=K_n+1}^{\infty} \mathbb{P}\Big(K_n d^+(\hat{\mu}_{a,s},\mu_1) < \log(n) + 3\log(\log(n))\Big)$$

$$= K_n + \sum_{s=K_n+1}^{\infty} \mathbb{P}\left(d^+(\hat{\mu}_{a,s},\mu_1) < \frac{d(\mu_a,\mu_1)}{1+\epsilon}\right)$$

$$\leq \frac{1+\epsilon}{d^+(\mu_a,\mu_1)}\Big(\log(n) + 3\log(\log(n))\Big) + \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}}$$

■

according to Lemma 4.

Lemma 3

$$\sum_{t=1}^{n} \mathbb{1}\{A_t = a, \mu_1 \leq u_1(t)\} \leq \sum_{s=1}^{n} \mathbb{1}\{sd^+(\hat{\mu}_{a,s},\mu_1) < \log(n) + 3\log(\log(n))\}\ .$$

Proof Observe that $A_t = a$ and $\mu_1 \leq u_1(t)$ together imply that $u_a(t) \geq u_1(t) \geq \mu_1$, and hence that

$$d^+\left(\hat{\mu}_a(t),\mu_1\right) \leq d(\hat{\mu}_a(t),u_a(t)) = \frac{\log(t) + 3\log(\log(t))}{N_a(t)}\ .$$

Thus,

$$\sum_{t=1}^{n} \mathbb{1}\{A_t = a, \mu_1 \leq u_1(t)\} \leq \sum_{t=1}^{n} \mathbb{1}\{A_t = a, N_a(t)d^+(\hat{\mu}_a(t),\mu_1) \leq \log(t) + 3\log(\log(t))\}$$

$$= \sum_{t=1}^{n}\sum_{s=1}^{t} \mathbb{1}\{N_t(a) = s, A_t = a, sd^+(\hat{\mu}_{a,s},\mu_1) \leq \log(t) + 3\log(\log(t))\}$$

$$\leq \sum_{t=1}^{n}\sum_{s=1}^{t} \mathbb{1}\{N_t(a) = s, A_t = a\}\mathbb{1}\{sd^+(\hat{\mu}_{a,s},\mu_1) \leq \log(n) + 3\log(\log(n))\}$$

$$= \sum_{s=1}^{n} \mathbb{1}\{sd^+(\hat{\mu}_{a,s},\mu_1) \leq \log(n) + 3\log(\log(n))\}\sum_{t=s}^{n} \mathbb{1}\{N_t(a) = s, A_t = a\}$$

$$= \sum_{s=1}^{n} \mathbb{1}\{sd^+(\hat{\mu}_{a,s},\mu_1) \leq \log(n) + 3\log(\log(n))\}\ ,$$

as, for every $s \in \{1,\ldots,n\}$, $\sum_{t=s}^{n} \mathbb{1}\{N_t(a) = s, A_t = a\} \leq 1$. ■

Lemma 4 *For each $\epsilon > 0$, there exist $C_2(\epsilon) > 0$ and $\beta(\epsilon) > 0$ such that*

$$\sum_{s=K_n+1}^{\infty} \mathbb{P}\left(d^+(\hat{\mu}_{a,s},\mu_1) < \frac{d(\mu_a,\mu_1)}{1+\epsilon}\right) \leq \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}}\ .$$

Proof If $d^+(\hat{\mu}_{a,s},\mu_1) < d(\mu_a,\mu_1)/(1+\epsilon)$, then $\hat{\mu}_{a,s} > r(\epsilon)$, where $r(\epsilon) \in]\mu_a,\mu_1[$ is such that $d(r(\epsilon),\mu_1) = d(\mu_a,\mu_1)/(1+\epsilon)$. Hence,

$$\mathbb{P}\left(d^+(\hat{\mu}_{a,s},\mu_1) < \frac{d(\mu_a,\mu_1)}{1+\epsilon}\right) \leq \mathbb{P}\left(d(\hat{\mu}_{a,s},\mu_a) > d(r(\epsilon),\mu_a),\ \hat{\mu}_{a,s} > \mu_a\right)$$

$$\leq \mathbb{P}(\hat{\mu}_{a,s} > r(\epsilon)) \leq \exp(-sd(r(\epsilon),\mu_a))\ ,$$

and

$$\sum_{s=K_n+1}^{\infty} \mathbb{P}\left(d^+(\hat{\mu}_{a,s},\mu_1) < \frac{d(\mu_a,\mu_1)}{1+\epsilon}\right) \leq \frac{\exp(-d(r(\epsilon),\mu_a)K_n)}{1-\exp(-d(r(\epsilon),\mu_a))} \leq \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}} ,$$

with $C_2(\epsilon) = (1-\exp(-d(r(\epsilon),\mu_a)))^{-1}$ and $\beta(\epsilon) = (1+\epsilon)d(r(\epsilon),\mu_a)/d(\mu_a,\mu_1)$. Easy computations show that $r(\epsilon) = \mu_a + O(\epsilon)$, so that $C_2(\epsilon) = O(\epsilon^{-2})$ and $\beta(\epsilon) = O(\epsilon^2)$. ■

Theorem 5 *Let $(X_t)_t \geq 1$ be a sequence of independent random variables bounded in $[0,1]$ defined on a probability space $(\Omega,\mathcal{F},\mathbb{P})$ with common expectation $\mu = \mathbb{E}[X_t]$. Let $\mathcal{F}_t$ be an increasing sequence of σ-fields of $\mathcal{F}$ such that for each t, $\sigma(X_1\ldots,X_t) \subset \mathcal{F}_t$ and for $s > t$, X_s is independent from $\mathcal{F}_t$. Consider a previsible sequence $(\epsilon_t)_{t\geq 1}$ of Bernoulli variables (for all $t > 0$, ϵ_t is $\mathcal{F}_{t-1}$-measurable). Let $\delta > 0$ and for every $t \in \{1,\ldots,n\}$ let*

$$S(t) = \sum_{s=1}^{t} \epsilon_s X_s , \qquad N(t) = \sum_{s=1}^{t} \epsilon_s , \qquad \hat{\mu}(t) = \frac{S(t)}{N(t)} ,$$
$$u(n) = \max\left\{q > \hat{\mu}_n : N(n)d\left(\hat{\mu}(n),q\right) \leq \delta\right\} .$$

Then

$$\mathbb{P}\left(u(n) < \mu\right) \leq e\left\lceil \delta\log(n)\right\rceil \exp(-\delta) .$$

4. Conclusion

The main result of the paper is Theorem 2. KL-UCB is the first index policy that reaches the lower-bound of Lai and Robbins (1985) for binary rewards.

In experiments, KL-UCB performs well in scenario 1: two arms, scenario 2: low rewards and scenario 3: bounded exponential rewards.

In this approach, only an upper-bound of the deviations (more precisely, of the exponential moments) of the rewards is required, which makes it possible to obtain versatile policies satisfying interesting regret bounds for large classes of reward distributions. The resulting index policies are simple, fast, and very efficient in practice, even for small time horizons.

References

J-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Resaerch*, 11:2785–2836, 2010.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvari. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410:1876–1902, 2009.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Apostolos Burnetas and Michael N. Katehakis. Optimal adaptive policies for markov decision processes. *Math. Oper. Res.*, 22:222–255, 1997.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. 1985.