Kullback-Leibler Maillard Sampling for Multi-armed Bandits with Bounded Rewards

Hao Qin, Kwang-Sung Jun, Chicheng Zhang

THE UNIVERSITY OF ARIZONA

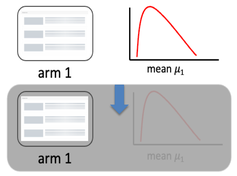
NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

We study multi-armed bandits with $[0,1]$-bounded reward distributions. We analyze Kullback-Leibler Maillard Sampling (KL-MS), a randomized exploration algorithm inspired by Minimum Empirical Divergence (MED) [Honda & Takemura, 2011] and Maillard Sampling (MS) [Maillard, 2013, Bian & Jun, 2022] that has closed form arm sampling probabilities, which enables unbiased offline evaluation. We give an analysis of KL-MS that yields both finite-time and asymptotic regret guarantees. In addition, we report an adaptive worst-case regret bound $O\left(\sqrt{\mu^*(1-\mu^*)KT\ln(K)}\right)$ for KL-MS. Such adaptivity is the first time reported in the literature.

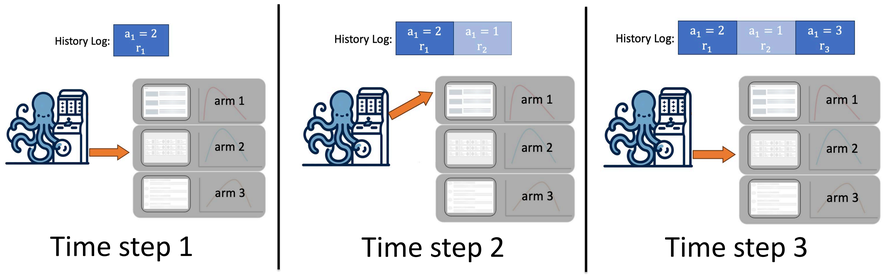
Multi-armed Bandits with [0,1]-Bounded reward

- Arm set $[1,\dots,K]$ with K arms in total, reward distributions $(\nu_a)_{a\in\mathcal{A}}$ supported in $[0,1]$ with mean vector $(\mu_a)_{a\in\mathcal{A}}$ and time length T.
- For time step $t = 1,\dots,T$,
  1. Agent pulls arm a_t.
  2. Return a reward $r_t \sim \nu_{a_t}$.



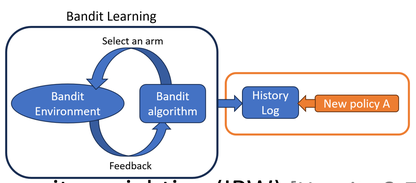
- Each time the agent suffers a suboptimality gap $\Delta_a := \mu^* - \mu_a$, μ^* is the optimal reward.
- Overall the agent need to minimize the pseudo-regret,

$$\textbf{Regret}(T) = \sum_{t=1}^{T} \Delta_{a_t} = \sum_{t=1}^{T} \mu^* - \mu_{a_t}$$



Offline Evaluation

- Evaluate a policy using the history log, instead of running in the real environment and causing excessive losses.



- Using Inverse propensity weighting (IPW) [Horvitz & Thompson, 1952] which is an unbiased estimator to evaluate the expected performance of the new policy:

Estimation target: μ_a | **IPW Estimator : $\hat{r}_t = \frac{I\{a_t=a\}r_t}{p_{t,a}}$**

$p_{t,a}$ is the probability of pulling arm a at time step t which can be part of the history log.

Performance Measures

- **Asymptotic Optimality**[Lai & Robbins, et al., 1985] in the Bernoulli reward setting

$$\limsup_{T\to\infty} \frac{\text{Regret}(T)}{\ln(T)} = \sum_{a:\Delta_a>0} \frac{\Delta_a}{\text{kl}(\mu_a,\mu^*)}$$

- **Minimax Ratio**[Audibert & Bubeck, et al., 2009][Auer et al., 2002]

An algorithm is said to have minimax ratio $f(K,T)$ if for any $[0,1]$-Bounded reward MAB environment,

$$\text{Regret}(T) \le O\left(\sqrt{KT}f(K,T)\right)$$

- **Sub-UCB** [Lattimore, 2018]:

$$\text{Regret}(T) \le O\left(\sum_{a:\Delta_a>0} \frac{\ln(T)}{\Delta_a}\right)$$

- **Closed-form Arm Sampling Distribution**

$p_{t,a}$ can be calculated in closed form, enabling unbiased estimation using the IPW estimator.

Kullback-Leibler Maillard Sampling

Each time step, the agent pulls arm following the **arm sampling probability distribution $p_{t,a}$**:

$$p_{t,a} \propto \exp\left(-N_{t-1,a} \cdot kl\left(\hat{\mu}_{t-1,a}, \hat{\mu}_{t-1,\max}\right)\right)$$

- $\text{kl}(\mu_1, \mu_2) := \mu_1 \ln\left(\frac{\mu_1}{\mu_2}\right) + (1-\mu_1)\ln\left(\frac{1-\mu_1}{1-\mu_2}\right)$ is the binary KL divergence.
- $N_{t-1,a}$ is the number of arm a has been pulled at time step $t-1$.
- $\hat{\mu}_{t-1,a}$ is the empirical mean of arm a at time step $t-1$.
- $\hat{\mu}_{t-1,\max}$ is the best empirical mean at time step $t-1$.

Comparison with Related Works

- **Minimum Empirical Discrepancy** [Honda & Takemura, 2011] : MED works in the finite bounded reward setting and its arm sampling rule is

$$p_{t,a} \propto \exp\left(-N_{t-1,a} \cdot \mathcal{K}_{\text{inf}}\left(\hat{\mu}_{t-1,\max}, \hat{\nu}_{t-1,a}, \hat{\nu}_{t-1,\max}\right)\right)$$

For each arm a, we aim to minimize the KL divergence to find the distribution closest to the empirical distribution $\hat{\nu}_{t-1,a}$. This distribution should have an expectation greater than $\hat{\mu}_{t-1,\max}$ and a support set defined by the historical returned rewards of arm a. We then evaluate $\mathcal{K}_{\text{inf}}$ as the KL divergence between this closest distribution and $\hat{\nu}_{t-1,\max}$. Only if when the reward distributions are Bernoulli, MED is equivalent to KL-MS.

- **Maillard Sampling** [Maillard, 2013][Bian & Jun, 2022] : MS works in the σ^2-sub-Gaussian reward setting and its sampling rule is

$$p_{t,a} \propto \exp\left(-N_{t-1,a} \cdot \frac{\left(\hat{\mu}_{t-1,a} - \hat{\mu}_{t-1,\max}\right)^2}{2\sigma^2}\right)$$

If Focusing on $[0,1]$-Bounded reward distribution, KL-MS seems more appropriate than MS.

Theoretical Comparison

Algorithm & Analysis	Asymptotic Optimality in Bernoulli	Minimax Ratio	Sub-UCB	Closed-form Arm Sampling Distribution
Thompson Sampling [Agrawal et al., 2013]	Yes	$\sqrt{\ln(K)}$	Yes	No
ExpTS [Jin et al., 2022]	Yes	$\sqrt{\ln(K)}$	Yes	No
ExpTS+ [Jin et al, 2022]	Yes	1	N/A**	No
UCB1 [Auer et al, 2002]	No	$\sqrt{\ln(T)}$	Yes	N/A (Deterministic)
KL-UCB++ [Menard & Garivier, 2017]	Yes	1	N/A**	N/A (Deterministic)
KL-UCB-Switch [Garivier et al., 2022]	Yes	1	N/A**	N/A (Deterministic)
MED [Honda & Takemura, 2011]	Yes	N/A	N/A	Yes
MS [Maillard, 2013] [Bian & Jun, 2022]	No	$\sqrt{\ln(T)}$	Yes	Yes
KL-MS	Yes	$\sqrt{\ln(K)}$	Yes	Yes

**: we conjecture that the answer is no.

Experimental Result

1. Regret Comparison:

KL-MS vs MS vs Bernoulli Thompson Sampling (BernoulliTS) in two 2-armed bandit instances borrow from [Kaufmann et al., 2012].

Figure 1. regret comparison $\mu = [0.8, 0.9]$

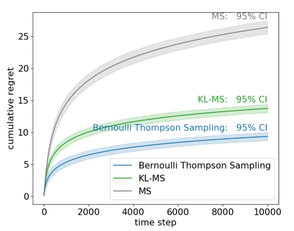
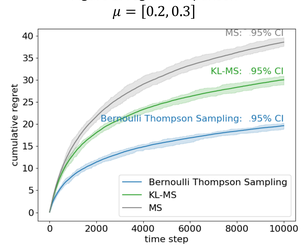


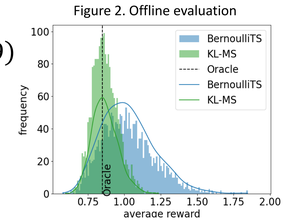
Figure 2. regret comparison $\mu = [0.2, 0.3]$



2. Offline Evaluation:

- 2-armed bandit instances, $(\mu_1, \mu_2) = (0.8, 0.9)$
- KL-MS records $p_{t,a}$ directly.
- BernoulliTS uses Monte Carlo method to estimate $p_{t,a}$ by $\hat{p}_{t,a}$
- Uniform policy:

Figure 2. Offline evaluation



Estimation target : $\mu = \frac{1}{K}\sum_{a:a\in[K]} \mu_a$

Pseudo-IPW Estimator : $\frac{1}{TK}\sum_{t=1}^{T} \frac{I\{a_t=a\}r_t}{\hat{p}_{t,a}}$ (BernoulliTS)

IPW Estimator : $\frac{1}{TK}\sum_{t=1}^{T} \frac{I\{a_t=a\}r_t}{p_{t,a}}$ (KL-MS)

- The log of KL-MS enables unbiased estimation.

SCAN ME