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

Hao Qin
University of Arizona
hqin@math.arizona.edu

Kwang-Sung Jun
University of Arizona
kjun@cs.arizona.edu

Chicheng Zhang
University of Arizona
chichengz@cs.arizona.edu

Bandit Problem

- Bandit problem have been applied in many online applications.

Online advertisement

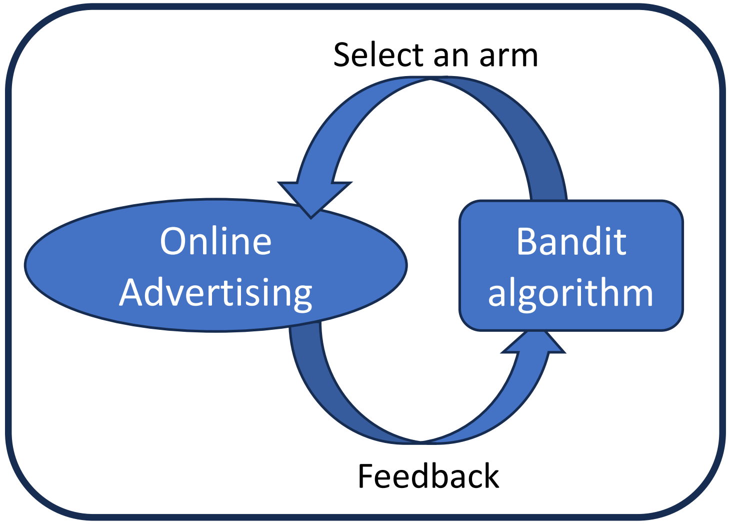
Layout A Layout B Layout C

How to select the most appealing website layout for online advertising?

Bandit Problem

- Offline policy evaluation

Real world environment

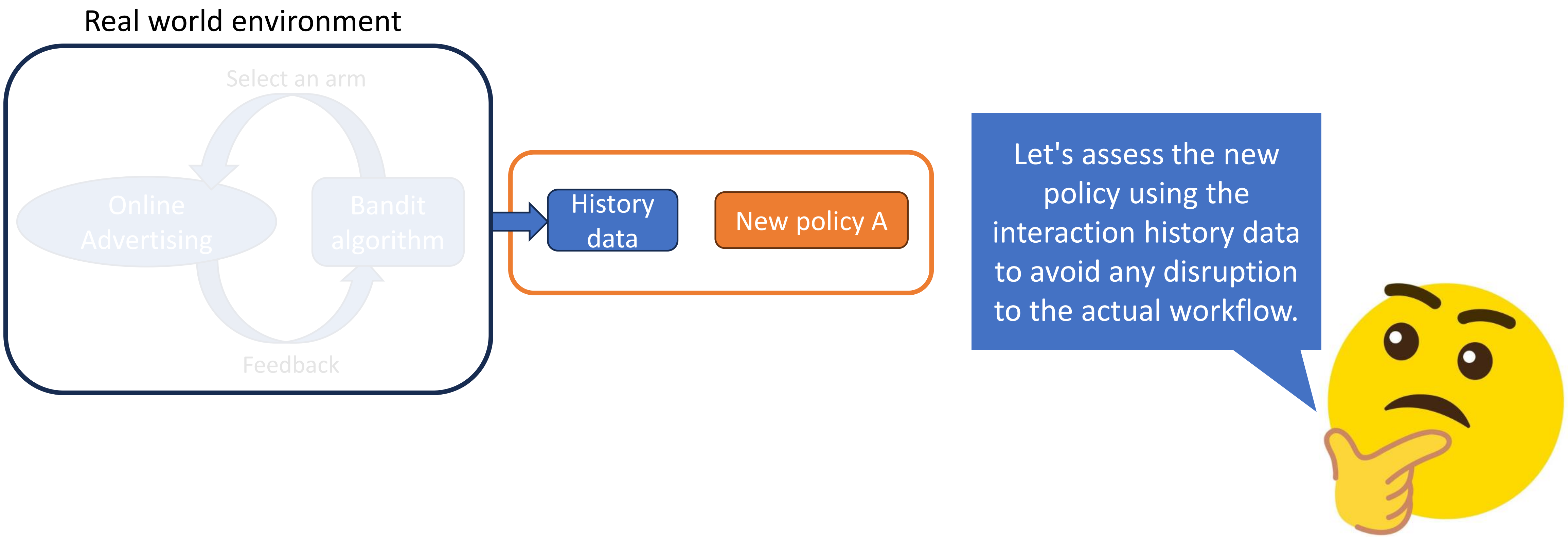


New policy A

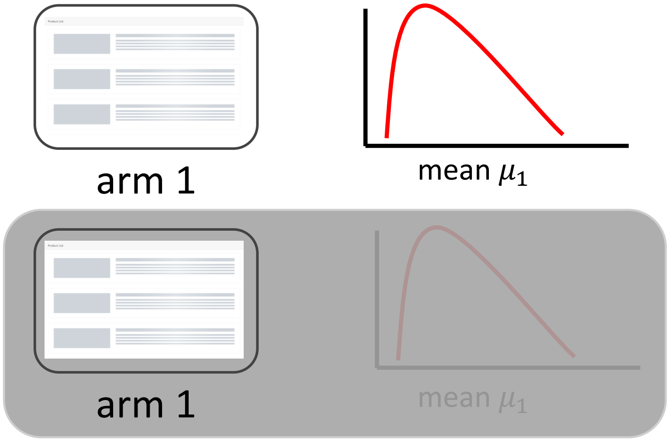
I'd like to evaluate the effectiveness of the new policy A without disrupting the actual workflow in a real-world environment

Bandit Problem

- Offline policy evaluation

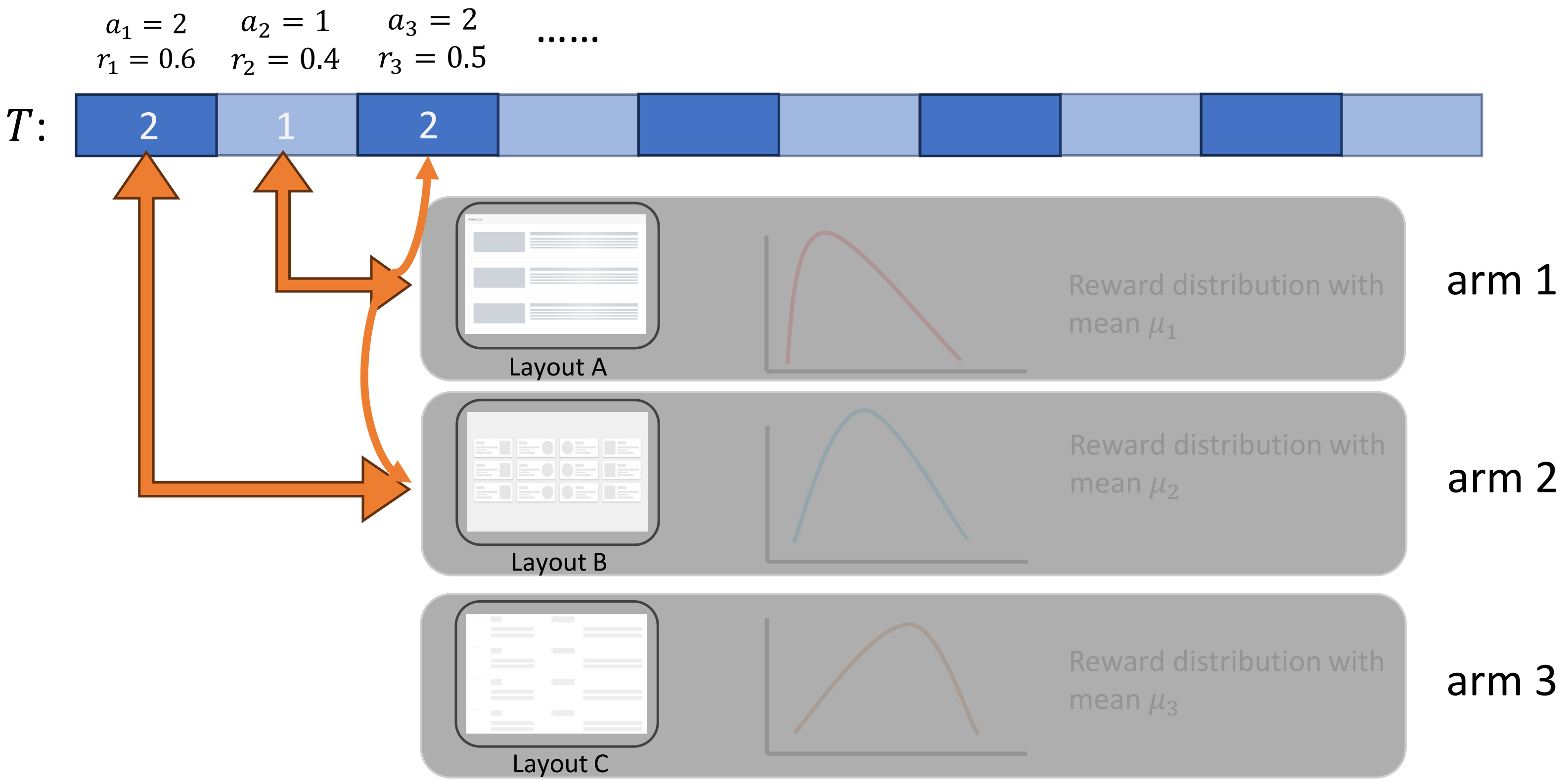


Problem setting



- K-arm bandits
- Each arm a is associated with a Bernoulli distribution with unknown mean $\mu_a \in [0,1]$

Problem setting



- K-arm bandits
- Each arm a is associated with a Bernoulli distribution with unknown mean $\mu_a \in [0,1]$

- T is the time horizon
- Record the arm pulling a_t
- Record the returned reward r_t
- Pull one arm and suffer a suboptimality:

$$\max_{a\in[K]} \mu_a - \mu_{a_t} =: \Delta_{a_t}$$

- $\mathbf{Regret(T)} = \sum_{t=1}^{T} \Delta_{a_t}$

Regret measurement

From an asymptotic perspective:

- Asymptotic optimality for Bernoulli Distribution[Lai & Robbins, 1985]:

$$\liminf_{T\to\infty} \frac{\mathrm{Regret}(T)}{\ln(T)} = \sum_{a:\Delta_a>0} \frac{\Delta_a}{kl(\mu_a,\mu^\star)},\ \mu^\star \text{is the optimal reward}$$

In a finite-time regime:

- Minimax Ratio for K-armed bandit [P. Auer et al., 2002][J. Y. Audibert et al., 2009]:

$$\frac{\mathrm{Regret(T)}}{\sqrt{KT}} = f(K,T),\ O(\sqrt{KT}) \text{ is the minimax optimal}$$

- Sub-UCB criteria [Lattimore, 2018]:

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

Prior works

Algorithm & Analysis	Asymptotic Optimality	Minimax Ratio	Sub-UCB	Closed-form Sampling dist.
TS [S. Agrawal et al, 2013] ExpTS [T. Jin et al, 2022]	Yes	$\sqrt{\ln(K)}$	Yes	No
ExpTS+ [T. Jin et al, 2022]	Yes	1	No	No
KL-UCB++ [P. Menard, A. Garivier, 2017] KL-UCB-Switch [A. Garivier et al., 2022]	Yes	1	N/A**	N/A (Deterministic)
MED [J. Honda, A. Takemura, 2011]	Yes	N/A	N/A	Yes
MS [B. Jie, K. Jun, 2021][O. A. Maillard, 2013]	No	$\sqrt{\ln(T)}$	Yes	Yes
KL-MS	Yes	$\sqrt{\ln(K)}$	Yes	Yes

**: we conjecture that the answer is no.

Algorithm design

The sampling probability distribution $p_{t,a}$:

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

- kl$(\mu_1,\ \mu_2)$ denotes the binary KL divergence to measure the distance between the sample arm and the empirical best arm. (kl$(\mu_1,\ \mu_2) \coloneqq \mu_1 \ln \frac{\mu_1}{\mu_2} + (1-\mu_1) \ln \frac{1-\mu_1}{1-\mu_2}$)
- $N_{t-1,a}$ is the number of arm a been pulled up to time $t-1$.
- $\hat{\mu}_{t-1,a}$ is the empirical mean of arm a up to time $t-1$.
- $\hat{\mu}_{t-1,\max}$ is the best empirical mean up to time $t-1$.

Conclusion

- KL-MS satisfies asymptotically optimality, better minimax ratio, and sub-UCB criterion.
- KL-MS has an adaptive worst-case regret bound $\sqrt{\mu^*(1-\mu^*)KT\ln(K)}$
- KL-MS has an unbiased estimation in the offline evaluation.

Thank You