An Optimal Algorithm for the Stochastic Bandits with Knowing Near-optimal Mean Reward

Extended Abstract

Shangdong Yang, Hao Wang, Yang Gao
State Key Laboratory for Novel Software Technology at
Nanjing University
Nanjing, China
yangshangdong007@gmail.com
{wanghao,gaoy}@nju.edu.cn

Xingguo Chen
School of Computer Science and Technology, School of
Software, Nanjing University of Posts and
Telecommunications
Nanjing, China
chenxg@njupt.edu.cn

ABSTRACT

This paper studies a variation of stochastic multi-armed bandit (MAB) problem where the agent knows a prior knowledge named Near-optimal Mean Reward (NoMR). We show that the cumulative regret of this bandit variation has a lower bound of $\Omega(1/\Delta)$, where Δ is the gap between the optimal and the second optimal mean reward. An algorithm called NoMR-Bandit is proposed to this variation, and we demonstrate that the cumulative regret of NoMR-Bandit has a uniform upper bound of $O(\Delta)$. It is concluded that NoMR-Bandit is optimal in terms of the order of regret bounds.

KEYWORDS

Multi-armed bandit problem; Learning theory; Reward structures for learning

ACM Reference Format:
Shangdong Yang, Hao Wang, Yang Gao and Xingguo Chen. 2018. An Optimal Algorithm for the Stochastic Bandits with Knowing Near-optimal Mean Reward. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

1 INTRODUCTION

The stochastic multi-armed bandit problem with knowing optimal mean reward has been studied in the last decays [2, 3, 5, 7]. In their settings, the agent only knows the exact optimal mean reward. But in more cases, the optimal mean reward cannot be exactly acquired. For example, in on-line web ad services, the mean reward of a recommended item is click through rate which is an important heuristic information to the recommender system [4]. Furthermore, the near-optimal item click through rate of a user can be associated with his/her demographic characteristics [1], but this click rate is an estimation. The recommendation engine can use this estimated rate as a prior knowledge to speed up learning.

There have been some algorithms to deal the bandits with knowing the prior knowledge. GCL* [7] and APT [5] are two deterministic algorithms, they take the optimal mean reward as the prior knowledge and the regret bound of them are logarithmically related to the time step. However, they can not handle the problem when the prior knowledge is less than the optimal mean reward. BPR [3] and BL [2] are two randomized algorithms, they take the optimal mean reward and the gap between the optimal and the second optimal mean reward as the prior knowledge. They achieve uniform upper bounds, but they also have the same problem with that of GCL* and APT. VZ [8] is a randomized algorithm, it takes the near-optimal mean reward as the prior knowledge and achieves a uniform upper bound. But its exploration is inefficient because it does not use the prior knowledge during exploration phase.

To achieve an efficient exploration and a uniform upper bound, this paper proposes a novel parameter-free stochastic bandit algorithm in the setting when the agent knows NoMR before. The key idea of our method (named NoMR-Bandit) is that we design a bimodal function with NoMR during exploration phase to encourage an accurate exploration. Specifically, instead of uniform exploration, NoMR-Bandit prefers to explore those arms whose estimated mean reward are small or close to NoMR. We give the lower bound of the bandit problem with NoMR ($\Omega(1/\Delta)$) and the associated upper bound of NoMR-Bandit ($O(\Delta)$).

2 PROBLEM FORMULATION

2.1 Stochastic Bandits with Knowing NoMR

In stochastic multi-armed bandit problem, the reward of each arm $i \in \{1, 2, ..., K\}$ in the arm set $\mathcal{A}$ corresponds to an unknown probability distribution ν_i. At each time step $t = 1, 2, ...n$, the agent selects an arm I_t from the arm set and receives a reward r_{I_t} drawn from ν_{I_t} (independent from the past). Denote by μ_i the mean reward of arm i and define

$$\mu^* = \max_{i\in\mathcal{A}} \mu_i \text{ and } i^* \in \operatorname*{argmax}_{i\in\mathcal{A}} \mu_i.$$

Without loss of generality, we assume that for each arm $i \in \mathcal{A}$, $\mu_i \in [0, 1]$. Let $T_i(n) = \sum_{t=1}^{n} \mathbb{1}_{I_t=i}$ denote the number of times the agent selected the arm i in the n rounds. The estimated $\hat{\mu}_i$ of each arm at round n is defined as

$$\hat{\mu}_i = \frac{1}{T_i(n)} \cdot \sum_{t=1}^{n} r_{I_t} \cdot \mathbb{1}_{I_t=i}.$$

Without loss of generality, during the learning process, we assume the largest estimated mean reward is $\hat{\mu}_1$.

Definition (NoMR). The near-optimal mean reward η is a real value between the optimal and the second optimal mean reward. η is defined as

$$\mu^* - \Delta < \eta \leq \mu^*.$$

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

Here we define the gap δ as the difference value between the optimal mean reward μ^* and the near-optimal mean reward η,

$$\delta = \mu^* - \eta.$$

And with the definition of η, $\delta \in [0, \Delta)$.

Objective Let $\Delta_i = \mu^* - \mu_i$ denote the gap between the optimal arm i^* and arm i, and let $\Delta = \min_i \Delta_i$. The objective in this bandit setting is to design an arm choosing policy to minimize the cumulative regret which is defined as

$$R_n = n\mu^* - \mathbb{E}\sum_{t=1}^{n} \mu_{I_t} = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)],$$

where $\mathbb{E}[T_i(n)]$ is the expectation of $T_i(n)$.

2.2 Lower bound

This result is analyzed based on a simple two-armed case where the reward distribution is assumed as normal distribution because it remains the same order for all families of distributions [6].

Here we denote by $\nu = \nu_1 \otimes \nu_2$ the product distribution that generates the rewards from ν_j when pulling arm $j \in \{1, 2\}$.

Theorem 2.1. *Let $\nu_0 = \mathcal{N}(\eta - \Delta_0, 1) \otimes \mathcal{N}(\eta + \Delta_1, 1)$ and $\nu_1 = \mathcal{N}(\eta + \Delta_1, 1) \otimes \mathcal{N}(\eta - \Delta_0, 1)$ where $\Delta_0 \in [0, \eta]$ and $\Delta_1 \in [0, 1 - \eta]$. Then for any policy and any $n \geq 1$, the lower bound of R_n of proposed setting can be divided into two cases.*

For $\delta \in (0, \Delta)$, which means $\mu^* - \Delta < \eta < \mu^*$,

$$\max(R_n(\nu_0), R_n(\nu_1)) \geq \frac{1}{4(\Delta_0 + \Delta_1)}.$$

For $\delta = 0$, which means $\eta = \mu^*$, there exists $c \in (0, 1]$ and a bandit problem with the smallest gap Δ and the known optimal mean reward,

$$R_n \geq \frac{\log(n/139)}{2\Delta}.$$

3 OUR METHOD

3.1 NoMR-Bandit

NoMR-Bandit receives the prior knowledge η and the number of arms K of the bandit problem as input. First, the agent pulls each arm of the game once. At time $t > K$, NoMR-Bandit checks that if the arm set $\mathcal{S}_t = \{i : \hat{\mu}_i \geq \eta\}$ is empty. If it is not empty, then the agent takes *Exploitation* strategy, it pulls the arm with the largest estimated mean reward in $\mathcal{S}_t$; otherwise, it makes *Exploration* according to the distribution in Algorithm 1.

It is important to define a a bimodal function in the exploration of NoMR-Bandit

$$\psi(x) = \left(x \cdot \frac{x - \eta}{\eta}\right)^2 + C,$$

where C is an adaptive real value which ensures $\sum_{i=1}^{K} p_{i,t} = 1$ at time step t. This is a differentiable function and it has a fantastic property that it will be smaller when the variable x is close to 0 or η. This makes the exploration distribution different from the previous for the reason that it encourages both the arms with small estimated mean reward and those estimated mean rewards are close to the prior knowledge η. This exploration strategy is more reasonable than those strategies with uniform exploration or just exploring the arms with small estimated mean reward.

Algorithm 1: NoMR-Bandit

```
Input: NoMR $\eta$, number of arms $K$
At the first $K$ round:
Choose each arm and, receive the reward $r_{I_t}$
for $t = K + 1, \ldots, n$ do
    Define the arm set $\mathcal{S}_t = \{i : \hat{\mu}_i \geq \eta\}$
    if $\mathcal{S}_t \neq \varnothing$ then
        Pull arm $I_t = i$ with the largest $\hat{\mu}_i$ in $\mathcal{S}_t$
    else
        Select $I_t$ randomly an arm according to
        $p_{i,t} = \frac{c}{\psi(\hat{\mu}_i)}$, where
        $c = \sum_{j=1}^{K} \frac{1}{\psi(\hat{\mu}_j)}$, $\psi(\hat{\mu}_i) = \left(\hat{\mu}_i \cdot \frac{\hat{\mu}_i - \eta}{\eta}\right)^2 + C$
    Observe reward $r_{I_t} \sim \nu_{I_t}$ and update $\hat{\mu}_i$
Output: The optimal arm $i^* := \text{argmax}_{i \in \mathcal{A}} \hat{\mu}_i$
```

3.2 Upper bound

Theorem 3.1. *For $K > 0$ and $n > K$ and let the priori knowledge η satisfies $\mu^* - \Delta < \eta \leq \mu^*$. Under the fact that $\psi(x)$ is a differentiable function with an infimum $\inf_x \psi(x) = C$. The upper bound of R_n of NoMR-Bandit can also be divided into two cases.*

For $0 < \delta < \Delta$, R_n is bounded by

$$(K - 1)\frac{\eta^2}{2} + \sum_{i \neq i^*} \left\{\Delta_i + \frac{\Delta_i}{C}\left[\frac{4\psi(\delta)}{\delta^2} + \int_{\delta}^{\infty} \frac{2\psi'(x)}{e^{\frac{x^2}{2}} - 1} dx\right]\right\},$$

For $\delta = 0$, R_n is bounded by

$$(K - 1)\frac{\eta^2}{2} + \sum_{i \neq i^*} \left\{\Delta_i + \frac{\Delta_i}{C} \sum_{t=1}^{n} \mathbb{E}\psi(\hat{\mu}_1)\right\}.$$

Specifically, if $\psi(x)$ defined as $\psi(x) = x^2 \cdot \left(\frac{x - \eta}{\eta}\right)^2$, the upper bound can be further concluded as following.

For $0 < \delta < \Delta$, R_n is bounded by

$$(K - 1)\frac{\eta^2}{2} + \omega \sum_{i \neq i^*} \Delta_i,$$

where $\omega = 1 + \frac{1}{C \cdot \eta^2}\left[(\delta - \eta)^2 + 4e^{-\delta^2}(\delta^2 + 1)\right]$.

For $\delta = 0$, R_n is bounded by

$$(K - 1)\frac{\eta^2}{2} + \sum_{i \neq i^*} \left\{\Delta_i + \frac{v \cdot \Delta_i}{C} \log(n)\right\},$$

where $v = \mathbb{E}\left(r_{I_1 = i^*}\right)^2$.

4 CONCLUSION

This paper studies the bandit problem where the agent knows a prior knowledge named NoMR. First, we show the lower bound of this bandit setting and then a novel algorithm NoMR-Bandit is proposed with a uniform upper bound. It is concluded that NoMR-Bandit is optimal in terms of the order of regret bounds.

ACKNOWLEDGMENTS

This paper is supported by National NSFC (Grant No.61432008, 61503178, 61403208, 61175042, 61321491) and Natural Science Foundation of Jiangsu Province, China (Grant No. BK20150587).

REFERENCES

[1] Jöran Beel, Stefan Langer, Andreas Nürnberger, and Marcel Genzmehr. 2013. The Impact of Demographics (Age and Gender) and Other User-Characteristics on Evaluating Recommender Systems. In *Proc. of the 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013), September 22-26, 2013, Valletta, Malta*. Springer, 396–400.

[2] Sébastien Bubeck and Che-Yu Liu. 2013. Prior-free and Prior-dependent Regret Bounds for Thompson Sampling. In *Proc. of the 27th Annual Conference on Neural Information Processing Systems (NIPS 2013), December 5-8, 2013, Lake Tahoe, NV, USA*. MIT Press, 638–646.

[3] Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. 2013. Bounded Regret in Stochastic Multi-armed Bandits. In *Proc. of the 26th Annual Conference on Learning Theory, (COLT 2013), June 12-14, 2013, Princeton University, NJ, USA*. Springer, 122–134.

[4] Cheng Li, Yue Lu, Qiaozhu Mei, Dong Wang, and Sandeep Pandey. 2015. Click-through Prediction for Advertising in Twitter Timeline. In *Proc. of the 21st International Conference on Knowledge Discovery and Data Mining (SIGKDD 2015), August 10-13, 2015, Sydney, NSW, Australia*. ACM, 1959–1968.

[5] Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. 2016. An Optimal Algorithm for the Thresholding Bandit Problem. In *Proc. of the 33rd International Conference on Machine Learning (ICML 2016), June 19-24, 2016, New York City, NY, USA*. ACM, 1690–1698.

[6] Philippe Rigollet and Assaf Zeevi. 2010. Nonparametric Bandits with Covariates. In *Proc. of the 23rd Conference on Learning Theory (COLT 2010), June 27-29, 2010, Haifa, Israel*. Springer, 54–66.

[7] Antoine Salomon and Jean-Yves Audibert. 2011. Deviations of Stochastic Bandit Regret. In *Proc. of the 22nd International Conference on Algorithmic Learning Theory (ALT 2011), October 5-7, 2011, Espoo, Finland*. Springer, 159–173.

[8] Sattar Vakili and Qing Zhao. 2013. Achieving Complete Learning in Multi-Armed Bandit Problems. In *Proc. of the 47th Annual Asilomar Conference on Signals, Systems and Computers (ACSSC 2013), November 3-6, 2013, Pacific Grove, CA, USA*. IEEE, 1778–1782.