Bandit Problems in Wireless Communication Network

Yao Zhao
ShanghaiTech University
zhaoyao1@shanghaitech.edu.cn

Abstract

To achieve long-time system performance in the next generation wireless communication network, traditional optimization based method has been show incompetent to attain desirable effect. There are many works now focusing on some model-based method to cater the character of the next generation wireless communication network. Multi-armed bandit is one of the representative statistical learning model, which has shown power to handle resource management problem in wireless network. We investigate the taxonomy of Multi-armed bandit problems, and recent works combining these cases with wireless network problems. And we further discuss some potential opportunity to refine existing solutions.

1. Introduction

1.1. An introduction to Multi-armed bandit(MAB)

Multi-armed bandit problems are the most basic examples of sequential decision problems in statistical learning with an exploration-exploitation trade-off[21]. This is the balance between staying with the option that gave highest payoffs in the past and exploring new options that might give higher payoffs in the future. Although the study of bandit problems dates back to the 1930s, exploration-exploitation trade-off arise in several modern applications, such as ad placement, website optimization, and packet routing. Mathematically, a multi-armed bandit is defined by the payoff process associated with each option.

As a special case of online learning or online optimization[17], the algorithm for MAB problem usually be compared with an optimal solution named oracle where the complete knowledge of the underlying reward process is known. The MAB lectures usually use a concept named *regret* to evaluate the performance of the proposed algorithm. It was first introduced in[10], and now widely used in multi-armed bandit analysis[1]. Supposed there are $K \geq 2$ arms, the player takes an action from the corresponding action set S repeatedly within a given time horizon T. The regret is defined as follow:

$$R(T) = \sum_{t=1}^{T} \max_{i \in S} X(i,t) - \sum_{t=1}^{T} X(i^*,t)$$

where $X(-,-)$ donates the reward (usually be normalized), i^* is the action generated by the algorithm. Note that $R(T)$ is a non-decreasing function of T . For any algorithm to be able to learn effectively,

$R(T)$ has to grow sublinearly with T . In this way, one has $\lim_{T\to\infty} \frac{R(T)}{T} = 0$, indicating that asymptotically the algorithm has no performance loss against the genie-aided solution.

There are three fundamental formulations of the bandit problem depending on the assumed nature of the reward process[25]:

1. Stochastic bandit:

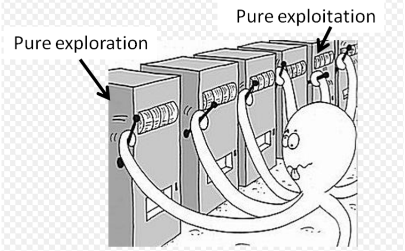
   The reward process of each arm is independent and identically distributed (i.i.d). It is often but not always assumed that the underlying distributions have a nite support. The rewards across different arms may be dependent.

2. Non-stochastic (adversarial) bandit:

   No stochastic assumption is made on the generation of rewards. One can view the rewards as being picked by an adversary. If the adversary generates the reward process ahead of the actual plays, such an adversary is called oblivious adversary as it is oblivious of the players strategy. On the other hand, at each time step, simultaneously with the players choice of the arm, a nonoblivious adversary assigns to each arm a reward. In the latter case, one can view the adversarial MAB problem as a game between the player and the adversary, and thus game theoretical results such as minmax regrets are often considered.

3. Markovian bandit:

   The reward process of each arm is modeled by a Markov chain with unknown transition probabilities and an unknown expected reward in each state. The oracle optimal policy for Markov MAB may alternate among multiple arms and can be determined using solutions to the corresponding Markov Decision Process (MDP).



There are two kinds of representative algorithm to solve MAB problems, the Upper-Condence Bound(UCB) algorithm for stochastic bandit problems and Exponential weights for Exploration and Exploitation(Exp3) randomized algorithm for adversarial bandit. The key idea behind these algorithm is to tracking the instantaneous expected reward by constructing some estimator that is usually not unbiased.

where:

$$UCB_{i,t} = \frac{\sum_{s=1,I_s=i}^{t} X(i,t)}{n_{i,t}} + \sqrt{\frac{\ln t}{n_{i,t}}}$$

and $n_{i,t}$ is the number of times arm i has been played up to time t. The second term in the right hand side is a bonus term. intuitively if there is an arm that is be pulled few times, this term would takes a large number to push the algorithm incline this arm.

```
Input: $K$ arms, number of rounds $T \geq K$
Output: action sequence
for $t = 1 \dots K$ do
    play arm $t$.
end
for $t = K \dots T$ do
    play arm:
        $$I_t = \arg\max_i UCB_{i,t}$$
end
```

Algorithm 1: UCB algorithm: simple vision

```
Input: $N$ arms, number of rounds $T \geq N$, a nonincreasing sequence of real numbers $\{\eta_t\}, t \in \mathbb{N}$,
        $p_1$ be the uniform distribution over all arms
Output: action sequence
for $t = 1 \dots T$ do
    Draw an arm $I_t$ from the probability distribution $p_t$;
    For each arm, compute the estimated loss $\hat{l_{i,t}} = \frac{l_{i,t}}{p_{i,t}} I_t = i$ and update the estimated cumulative
      loss $\hat{L_{i,t}} = \hat{L_{i,t-1}} + \hat{l_{i,t}}$;
    Compute the new probability distribution over arms according to:
        $$p_{i,t+1} = \frac{\exp(-\eta_t \widetilde{L}_{i,t})}{\sum_{k=1}^{K} \exp(-\eta_t \widetilde{L}_{k,t})}$$
end
```

Algorithm 2: Exp3 (Exponential weights for Exploration and Exploitation) algorithm

There are two fundamental ideas behind the Exp3 algorithm[3]: one is that despite the fact that only the loss of the played arm is observed, with a simple trick it is still possible to build an unbiased estimator for the loss of any other arm; another idea is to use an exponential reweighting of the cumulative estimated losses to dene the probability distribution from which the forecaster will select the arm.

1.2. Bandit problems in Wireless Communication

Due to the ever-increasing need for mobile services, we expect a massive growth in demand for wireless services in the years to come. As a result, future mobile networks are expected to accommodate new communications, networking, and energy-efficiency concepts, for instance, small cells, device-to-device (D2D) communications, and energy harvesting. In order to realize such novel concepts, system designers face some fundamental challenges. Any planning and scheduling in distributed networks depends heavily on the available energy at network nodes, energy levels are not observable in general. Therefore, it is evident that in future, efficient and robust wireless communications design needs to deal with inherent uncertainty and lack of information, as well as possible competition for resources, in a distributed manner. As a result, it becomes imperative to search for new mathematical tools to deal with networking problems, which in general involve both uncertainty and conflict[13].

In fact, MAB benefits from a wide range of variations in the setting, thus the family of MAB model has a big capacity to handle such problems.

2. Related Work

Perhaps the first work that uses MAB theorem to solve wireless Communication problem is [9]. This paper considers the design of medium access control protocols for cognitive radio networks in a highly dynamic environment. In the scenario under consideration, multiple cognitive users seek to opportunistically exploit the availability of empty frequency bands within parts of the radio spectrum having multiple bands. The availability of each channel is modelled as a Markov chain. The scenario in which the parameters of the Markov chain of each channel is known is first considered. An optimal symmetric strategy that maximizes the total throughput of the cognitive users is developed. Next, the situation in which the parameters of each channel are unknown a priori is considered. This problem is modelled as a competitive multiuser bandit problem. In order to cope with the phenomenal growth of mobile data trafc, unlicensed spectrum can be utilized by the Long Term Evolution (LTE) cellular systems. However, ensuring fair coexistence with WiFi is a mandatory requirement. In one approach, periodically congurable transmission gaps can be used to facilitate a coexistence between WiFi and LTE. In this paper, a Multi-Armed Bandit (MAB) based dynamic duty cycle selection method is proposed for conguration of transmission gaps ensuring a better coexistence for both technologies in[22]. To maximize cache utilization under the condition of varying and unknown popularity profile, [15] proposed a context-aware caching algorithm based on user information, which includes users ages, genders or their preferences. Algorithms for different variants of the contextual multi-armed bandit problem have been developed before, e.g.[12, 11, 23].

There are also fruitful works focusing on the mobility management(MM) in heterogeneous network (HetNet). The key challenge for efcient mobility management is the unavailability of accurate information of the candidate small base stations(SBSs) in an uncertain environment. The groundbreaking work is[20, 2]. It propose multi armed bandit and satisfaction based MM learning approaches aiming at improving the overall system performance and reducing the handover failure(HOF) and ping-pong(PP) probabilities. The main difference between wireless communication and theoretical MAB lectures is the latter usually do not consider the cost to switch among different arms. In fact, most of the stochastic bandit solutions incur fairly frequent exploration operations, which directly lead to the FHO problem, while for the real world application especially for the wireless communication problems the cost to switch between different base stations can not be ignored. [18] propose a non-stochastic online-learning approach, which does not make any assumption on the statistical behavior of the SBS activities. It allows the reward or cost when the user equipment(UE) choose one of the available SBSs to be any arbitrary sequence for time slot t. [19]is a representative method to solve FHO problem by introducing a increasing switching batch size to significantly reduce the number of handover. The most recent work is[24], it consider the similarity between mobility patterns to further explore latent progress in the 5G era.

3. Future work

There are still many drawbacks after talking about the above works. Due to the more and more complicated practical environment, dynamic SBSs situation need to be considered. For example, with the development of green communication, a new architecture using green base station has been proposed[16]. In addition to minimizing the environmental impact of the industry, cellular network operators are as well interested in reducing the energy consumption of their networks for economical reasons. The costs for running a network are largely affected by the energy bill and significant savings in capex and opex can be realized through reduced energy needs. From an economical perspective, we want to reduce the number of active SBSs, while maintain the total system performance. To handle this dynamic scene, the sleeping bandit[8, 6] model would be likely to be an useful mathematical tool. [18]consider a simple case of dynamic environment, where the SBSs could be turned on and off by users. In particular, such on/off behaivor would disrupt the learning process. In addition, since there is obvious user needs stratification, for example, some UEs may be more sensitive to the rate, other UEs may be more sensitive to the delay, a more efficient mobility management in the next generation network is to design a contextual and hierarchical model. [14] shows similar idea to handle a mobile crowdsourcing problem. Recently, contextual bandit problem has draw wide attention in machine learning society[7, 4, 5].

References

[1] S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[2] V. Capdevielle, A. Feki, and E. Sorsy. Joint interference management and handover optimization in lte small cells network. In *2012 IEEE International Conference on Communications (ICC)*, pages 6769–6773, June 2012.

[3] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[4] A. A. Deshmukh, U. Dogan, and C. Scott. Multi-task learning for contextual bandits. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4851–4859. Curran Associates, Inc., 2017.

[5] K. Greenewald, A. Tewari, S. Murphy, and P. Klasnja. Action centered contextual bandits. In *Advances in neural information processing systems*, pages 5979–5987, 2017.

[6] V. Kanade, H. B. McMahan, and B. Bryan. Sleeping experts and bandits with stochastic action availability and adversarial rewards. In *Artificial Intelligence and Statistics*, pages 272–279, 2009.

[7] A. Kazerouni, M. Ghavamzadeh, Y. Abbasi, and B. Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 3913–3922, 2017.

[8] R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma. Regret bounds for sleeping experts and bandits. *Machine learning*, 80(2-3):245–272, 2010.

[9] L. Lai, H. Jiang, and H. V. Poor. Medium access in cognitive radio networks: A competitive multi-armed bandit framework. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 98–102, Oct 2008.

[10] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[11] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[12] T. Lu, D. Pál, and M. Pál. Contextual multi-armed bandits. In *Proceedings of the Thirteenth international conference on Artificial Intelligence and Statistics*, pages 485–492, 2010.

[13] S. Maghsudi and E. Hossain. Multi-armed bandits with application to 5g small cells. *IEEE Wireless Communications*, 23(3):64–73, June 2016.

[14] S. Müller, C. Tekin, M. van der Schaar, and A. Klein. Context-aware hierarchical online learning for performance maximization in mobile crowdsourcing. *arXiv preprint arXiv:1705.03822*, 2017.

[15] S. Mller, O. Atan, M. van der Schaar, and A. Klein. Smart caching in wireless small cell networks via contextual multi-armed bandits. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–7, May 2016.

[16] L. Saker, S. E. Elayoubi, and T. Chahed. Minimizing energy consumption via sleep mode in green base station. In *2010 IEEE Wireless Communication and Networking Conference*, pages 1–6, April 2010.

[17] S. Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[18] C. Shen, C. Tekin, and M. van der Schaar. A non-stochastic learning approach to energy efficient mobility management. *IEEE Journal on Selected Areas in Communications*, 34(12):3854–3868, Dec 2016.

[19] C. Shen and M. van der Schaar. A learning approach to frequent handover mitigations in 3gpp mobility protocols. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, March 2017.

[20] M. Simsek, M. Bennis, and . Gven. Context-aware mobility management in hetnets: A reinforcement learning approach. In *2015 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1536–1541, March 2015.

[21] A. Slivkins. Introduction to multi-armed bandits.

[22] M. G. S. Sriyananda, I. Parvez, I. Gvene, M. Bennis, and A. I. Sarwat. Multi-armed bandit for lte-u and wifi coexistence in unlicensed bands. In *2016 IEEE Wireless Communications and Networking Conference*, pages 1–6, April 2016.

[23] C. Tekin and M. van der Schaar. Distributed online learning via cooperative contextual bandits. *IEEE transactions on signal processing*, 63(14):3700–3714, 2015.

[24] Z. Wang, Y. Xu, L. Li, H. Tian, and S. Cui. Handover control in wireless systems via asynchronous multi-user deep reinforcement learning. *arXiv preprint arXiv:1801.02077*, 2018.

[25] R. Zheng and C. Hua. *Sequential Learning and Decision-Making in Wireless Resource Management*. Springer, 2017.