Time-Decaying Bandits for Non-stationary Systems

Junpei Komiyama[1,⋆] and Tao Qin[2]

[1] The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
junpei@komiyama.info
[2] Microsoft Research, No. 5 Danling Street, Haidian District, Beijing, 100080, P.R. China
taoqin@microsoft.com

Abstract. Contents displayed on web portals (e.g., news articles at Yahoo.com) are usually adaptively selected from a dynamic set of candidate items, and the attractiveness of each item decays over time. The goal of those websites is to maximize the engagement of users (usually measured by their clicks) on the selected items. We formulate this kind of applications as a new variant of bandit problems where new arms are dynamically added into the candidate set and the expected reward of each arm decays as the round proceeds. For this new problem, a direct application of the algorithms designed for stochastic MAB (e.g., UCB) will lead to over-estimation of the rewards of old arms, and thus cause a misidentification of the optimal arm. To tackle this challenge, we propose a new algorithm that can adaptively estimate the temporal dynamics in the rewards of the arms, and effectively identify the best arm at a given time point on this basis. When the temporal dynamics are represented by a set of features, the proposed algorithm is able to enjoy a sub-linear regret. Our experiments verify the effectiveness of the proposed algorithm.

1 Introduction

The multi-armed bandit (MAB) problem is a typical example of sequential decision-making problems under uncertain environments and can model many real-world applications such as an adaptive routing, clinical trials, and a variety of recommendation problems. Among those applications, the recommendation problems, such as news recommendation [10] and social bookmarks [12], are attracting more and more attention from both the academia and the industry. Using the language of MABs, a recommendation problem can be described as follows. Given a set of K arms (candidate items) to select (display), in each round (user visit) the system selects one arm from the set and show it to the user. The system then receives a reward (whether the user clicks on the item or not) for the arm. The goal of MAB is to design an algorithm that optimizes the cumulative reward (the total number of user clicks given a number of user visits), which is achieved by accurately identifying the arm with the best expected reward (click-through rate (CTR)).

A well studied MAB problem is the so-called stochastic MAB, in which it is assumed that the rewards of each arm is i.i.d. drawn from a fixed but unknown distribution. The upper confidence bound (UCB) algorithm [6] is the standard algorithm in this problem.

⋆ This work was done while the author was visiting Microsoft Research Asia.

T.-Y. Liu et al. (Eds.): WINE 2014, LNCS 8877, pp. 460–466, 2014.

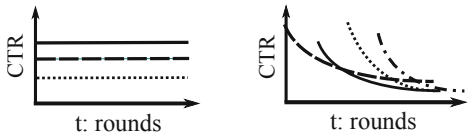

Fig. 1. Comparison between stochastic bandits and time-decaying bandits. The left figure shows a stochastic bandit where CTRs of the arms do not change over time. The right one shows a time-decaying bandit, where CTRs decay over rounds.

While the stochastic MAB successfully models many problems, it does not match well with the recommendation problems under our investigation, because it ignores an important factor of recommender systems: the attractiveness of an item to users decays over time. For example, it has been reported that new items (e.g., news articles in a web portals, and tweets in social networks) usually have larger CTRs than old ones [12], and a specific content will lose its attractiveness after being repeatedly displayed to the users [3]. In this situation, a simple application of the algorithms designed for stochastic MAB (e.g., UCB) will lead to over-estimation of the rewards of old arms, and thus cause a misidentification of the optimal arm.

Actually, the recommendation problems correspond to a new type of MAB problems, which we call the "time-decaying MAB", where the expected reward of each arm decays with respect to time (see Fig. 1). Before investigating such problems, the very first step is to characterize the decay factor. A simple way is to adopt a constant decay factor and integrate it into the stochastic MAB algorithms. However, in many real applications, such a characterization is inaccurate. This is because the decay factor (including magnitude and decreasing speed) varies largely among arms. Take news articles as examples. A breaking news usually quickly attracts a lot of attention, but the attention may drop significantly in just several hours. In contrast, news articles such as an enforcement of some national law usually warm up relatively slowly, but may attract a long-term attention like a week or so. In this case, it would be more appropriate to assume that the decaying factors for individual arms differ from each other and to learn them in an online manner. For this purpose, the learning algorithm needs to make exploration to simultaneously understand both the stochastic properties and the decaying factors of the rewards.

In this work, to solve the time-decaying MAB problems, we generalize the UCB algorithm by incorporating the information about the temporal dynamics into the computation of the upper confidence bounds of the arms. In particular, we represent the temporal dynamics by a set of basic time-dependent functions which consists of both fast and slow decays. The weights of individual functions are optimized by an application of the linear bandit technology. As a result, our algorithm is able to estimate the decay factor of each arm, which leads to an accurate estimation of the expected reward. Also, by choosing the arm of the largest index like the UCB algorithm, our algorithm balances exploration (give more chances to less selected arms for best arm identification) and exploitation (choosing the arm with the largest observed reward).

To summarize, the major contributions of our work lie in two aspects: (1) According to our knowledge, this is the first work that embeds temporal decay of the rewards into MAB problems. (2) We design an algorithm to solve the time-decaying MAB problems, the effectiveness of which is verified both theoretically and empirically.

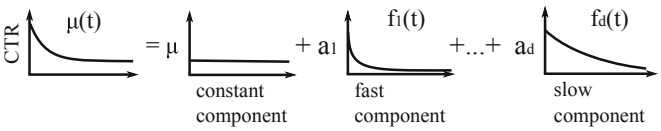

Fig. 2. CTR of an arm

2 Problem Setup

Time-decaying MAB extends stochastic MAB from two aspects: (1) new arms are continuously added to the candidate set, and (2) the expected reward of each arm decays in rounds. Let n be the total number of rounds, and $\mathcal{K}_t$ be the set of available arms in round $t \in [n]$. Let $r_{i,t}$ be the reward of arm i if it is selected at round t. In this work we take news recommendation as an example and exclusively consider the case of click-through feedback. Thus, in this case, $r_{i,t}$ is either 0 or 1, and the CTR of a news article corresponds to the expected reward of an arm. Henceforth, we use "news article" and "arm" interchangeably.

A time-decaying MAB goes as follows. At each round $t = 1, 2, ..., n$, a system selects one arm I_t from the candidate set $\mathcal{K}_t$ and receives a random reward $r_{I_t,t}$. Note that the reward information of the other arms are not available. The reward of arm i is drawn from a Bernoulli distribution parameterized by $\mu_{i,t}$, which assumes to be represented as the sum of a constant part μ_i and several basic decaying functions (see Fig. 2). Let $f_1(t-t_i), ..., f_d(t-t_i)$ be the set of basic decaying functions and $t-t_i$ be the number of rounds since arm i appears for the first time. The expected reward of arm i at round t can be modeled as $\mu_{i,t} = \mu_i + \sum_{k=1}^{d} a_{i,k} f_k(t-t_i)$, where $a_{i,k}$ is the weight associated with the k-th decaying function for arm i. Equivalently, by defining $\boldsymbol{x}_{i,t} = (1, f_1(t-t_i), \cdots, f_d(t-t_i)) \in R^{d+1}$ and $\boldsymbol{\theta}_i = (\mu_i, a_{i,1}, a_{i,2}, \cdots, a_{i,d}) \in R^{d+1}$, we can write $\mu_{i,t} = \boldsymbol{x}_{i,t}^\top \boldsymbol{\theta}_i$. That is, $\mu_{i,t}$ can be represented as a linear combination of a $(d+1)$-dimensional "context" $\boldsymbol{x}_{i,t}$ which consists of the constant 1 and the values of functions $f_1(t-t_i), ..., f_d(t-t_i)$. We assume that the contexts and the weights are bounded as $||\boldsymbol{x}_{i,t}|| \leq L$ ($||\boldsymbol{x}|| = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$) and $||\boldsymbol{\theta}_i|| \leq S$, respectively.

We define the optimal arm $i^*(t)$ as the arm with the largest expected reward at round t: $i^*(t) = \arg\max_{i \in \mathcal{K}_t} \mu_{i,t}$. Unlike the stochastic MAB, the optimal arm in our problem may vary in rounds, which is the essential difficulty in this problem. The performance of an algorithm is measured by the (pseudo) regret $\mathrm{R}(n)$, which is defined as the difference between the cumulative expected reward of the optimal pulling policy (which knows the expected rewards of all arms at each round) and that of the arms selected by the algorithm: $\mathrm{R}(n) = \sum_{t=1}^{n} \mu_{i^*(t),t} - \sum_{t=1}^{n} \mu_{I_t,t}$.

3 Algorithm

The key to solve the time-decaying MAB problem, where the reward of an arm can be represented as a linear combination of decaying functions, is to effectively estimate $\boldsymbol{\theta}_i$, the weights of individual decaying functions, for each arm. This falls into the framework

Algorithm 1. Time-decaying UCB

Inputs: $a(t), f_1(t-t_i), ..., f_d(t-t_i)$.
for $t = 1, 2, 3, ..., n$ **do**
  for $i \in \mathcal{K}_t$ **do**
    if arm i is new **then**
      $t_i = t$, $\mathbf{A}_{i,t} \leftarrow \mathbf{I}_{d+1}$ and $\boldsymbol{b}_{i,t} \leftarrow \mathbf{0}_{(d+1)\times 1}$
    end if
    $\boldsymbol{x}_{i,t} \leftarrow (1, f_1(t-t_i), ..., f_d(t-t_i))^T$
    $c_{i,t} \leftarrow a(t-t_i+1)||\boldsymbol{x}_{i,t}||_{\mathbf{A}_{i,t}^{-1}}$, and $\hat{\mu}_{i,t} \leftarrow \boldsymbol{x}_{i,t}^\top \mathbf{A}_{i,t}^{-1} \boldsymbol{b}_{i,t}$
    $g_{i,t} \leftarrow \hat{\mu}_{i,t} + c_{i,t}$
  end for
  Choose arm $I_t = \arg\max_{i \in \mathcal{K}_t} g_{i,t}$, and receive reward $r_{I_t,t} \in \{0, 1\}$
  $\mathbf{A}_{I_t,t+1} \leftarrow \mathbf{A}_{I_t,t} + \boldsymbol{x}_{I_t,t}\boldsymbol{x}_{I_t,t}^\top$ and $\boldsymbol{b}_{I_t,t+1} \leftarrow \boldsymbol{b}_{I_t,t} + r_{I_t,t}\boldsymbol{x}_{I_t,t}$
  for $i \neq I_t \in \mathcal{K}_t$ **do**
    $\mathbf{A}_{i,t+1} \leftarrow \mathbf{A}_{i,t}$ and $\boldsymbol{b}_{i,t+1} \leftarrow \boldsymbol{b}_{i,t}$
  end for
end for

of linear bandits [2,5,9,1], which perform an online estimation of linear weights with bandit feedback. The difference is that our problem contains multiple linear bandits: each arm can be considered as an instance of a linear bandit problem, whose context consists of a constant term and a series of temporal functions, while there is only one linear bandit in classical linear bandit problems[1]. In this sense, our problem is a hybrid of multi-armed bandits and linear bandits plus temporal decays.

Our proposed Algorithm 1 is shown as above, which we call time-decaying UCB. As can be seen, at each round of the algorithm, for each arm it constructs a matrix $\mathbf{A}_{i,t}$ and a vector $\boldsymbol{b}_{i,t}$, which are the sum of the covariance and the reward-weighted sum of features, respectively. $\hat{\mu}_{i,t}$, the least square estimation of the reward at round t, is given as $\boldsymbol{x}_{i,t}^\top \mathbf{A}_{i,t}^{-1} \boldsymbol{b}_{i,t}$. To guarantee the sufficient amount of exploration, we additionally introduce a confidence bound term $c_{i,t} = a(t)||\boldsymbol{x}_{i,t}||_{\mathbf{A}_{i,t}^{-1}}$, where $||\boldsymbol{x}||_{\mathbf{A}}$ is the matrix-induced vector norm $\sqrt{\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}}$. For the choice of $a(t)$, we show in the next section that a $O(\sqrt{\log t})$ function is appropriate to give a reliable confidence bound. Time-decaying UCB chooses the arm with the maximum UCB index $g_{i,t} = \hat{\mu}_{i,t} + c_{i,t}$.

4 Regret Bound

The following theorem shows that the proposed algorithm possesses a sublinear regret bound. The proof of the theorem, which combines the multi-armed bandit and the linear bandit techniques, is in the full version of this paper.

[1] There are several papers that analyze the regret of a linear bandit problem with multiple regressors (e.g., Cesa-Bianchi et al. [11], and Agrawal and Goyal [4]). However, The analysis in these papers are limited to the case where the set of arms does not change over time.

Theorem 1. *Let* $C(t,\delta') = \frac{1}{2}\sqrt{(d+1)\log\left(\frac{1+tL^2}{\delta'}\right)} + S$. *By setting* $\alpha(t) = C(t, \delta/|\mathcal{K}_t|)$, *the regret of the proposed algorithm is upper-bounded as follows with probability at least* $1-\delta$[2]

$$\mathrm{R}(n) \leq 2C(n,\delta/|\mathcal{K}_{all}|)\sqrt{|\mathcal{K}_{all}|n(d+1)(L^2+1)\log\left(1+\frac{nL^2}{d+1}\right)} = \tilde{O}(\sqrt{|\mathcal{K}_{all}|n}), \tag{1}$$

where $\tilde{O}$ *hides a polylog factor, and* $\mathcal{K}_{all}$ *is the set of all the arms through the run.*

5 Experiments

In this section, we report the results of our simulation. The goal here is to compare the empirical performance of the proposed algorithm with that of existing ones.

We simulate a news recommender system, where the decay pattern of CTR of each news article is different from the others and new articles are continuously added into the system.

Rounds and Articles: We try different values: $n = \{10^5, 10^{5\frac{1}{3}}, 10^{5\frac{2}{3}}, ..., 10^7\}$. For each n, all the algorithms are run for 20 times and the results are averaged over the runs. At the beginning of each run, there are 20 articles in the candidate set. Then new articles are continuously added into the set. At the end of each run, 100 articles are involved.

CTRs and Decay Factor: We set the CTR of the i-th article as $\mu_{i,t} = m_i y(\Delta_i(t)/t_{h,i})$, where m_i is its initial CTR when it is added into the system and $y(\Delta_i(t)/t_{h,i})$ is the decay function. The initial CTRs of all the news articles are independently drawn from a uniform distribution in $[0, 0.15]$. We adopt the square root decay function $y(x) = 1/\sqrt{x+1}$, which is reported and used in a contest of news article recommendation for Yahoo! Homepage [8]. $t_{h,i}$ defines how fast CTR decays, and is independently drawn from an exponential distribution $P(t_{h,i}) = -\lambda \exp(-t/\lambda)$ with $\lambda = 0.02n$. $\Delta_i(t) = t - t_i$ was the number of the rounds after the article is added into the system. In this setting, CTR of the optimal article averaged over time is around 0.1, which is similar to the case of the Yahoo! news article dataset [13]. Note that the above parameters (including the number of rounds, decay function, etc.) are notified to none of the algorithms. Fig. 3 displays a part of time series of CTRs of the articles in a run.

Compared Algorithms: We take RANDOM, Exp3.S [7] and UCB [6] as baselines for comparison: RANDOM is the algorithm that uniformly samples an article from all available ones; Exp3.S is a variant of the Exp3 algorithm for switching environment; and UCB is a stochastic bandit algorithm that ignores temporal dynamics. For our time-decaying UCB, we implement three variants with different sets of the temporal feature: (1) Decaying-UCB-3 has three temporal components $\{f_1(\Delta_i(t)), f_2(\Delta_i(t)), f_3(\Delta_i(t))\} = \{y(\Delta_i(t)/\lambda), y(2^4\Delta_i(t)/\lambda), y(2^8\Delta_i(t)/\lambda)\}$, (2) Decaying-UCB-5 has five temporal components $\{f_1(\Delta_i(t)), f_2(\Delta_i(t)), ..., f_5(\Delta_i(t))\} = \{y(\Delta_i(t)/\lambda), y(2^2\Delta_i(t)/\lambda), ..., y(2^8\Delta_i(t)/\lambda)\}$, and (3) Decaying-UCB-9 has nine

[2] We can set $\delta = O(1/n)$.

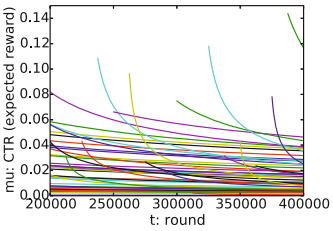

Fig. 3. CTRs (= expected rewards) of articles in a single run with $n = 10^6$. Each curve represents a CTR of an article. Most articles have low CTR, and the optimal article switches frequently.

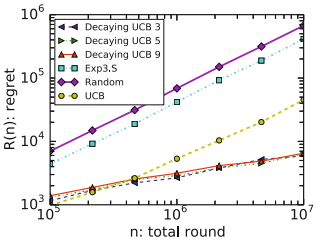

Fig. 4. The log-log plot of the regret of the algorithms. The horizontal axis is the number of total rounds (depend on the scale factor s) and the vertical line is the regret.

temporal components $\{f_1(\Delta_i(t)), f_2(\Delta_i(t)), ..., f_9(\Delta_i(t))\} = \{y(\Delta_i(t)/\lambda), y(2\Delta_i(t)/\lambda), ..., y(2^8\Delta_i(t)/\lambda)\}$. The hyper-parameters of the algorithms are set as follows: K in Exp3.S is set to 100, and S (switching number) is chosen best among $\{1, 2, 5, 10, 20, 50, 100\}$. $a(t)$ in UCB and in our time-decaying UCB, which determine the magnitude of the exploration, are chosen to be the best among $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0\} \times \sqrt{\log t}$.

Regret Comparison: Fig. 4 shows the regrets of all the algorithms. RANDOM and Exp3.S are clearly worse than UCB and time-decaying UCB. The performances of the three variants of time-decaying UCB are very close: our algorithm is not very sensitive to the selection of the basic decaying functions. Further, UCB and time-decaying UCB perform similarly when n is small, and the latter performs much better when $n \geq 10^6$. Considering that the number of daily visitors of a web portal spans from millions to hundreds of millions, it is rather easy to obtain a large n, and therefore our algorithm is expected to perform better than other algorithms in real-world recommender systems.

6 Conclusion

In this paper, we have proposed a new type of bandit problems in which new arms are dynamically added into the candidate set and the expected reward of each arm decays as the round proceeds. The performance of the proposed algorithm is verified both theoretically and empirically. For future works, we will (1) design new algorithms with better regret bounds and (2) consider more complicated dynamics in real-world applications that go beyond simple time decay.

References

1. Abbasi-Yadkori, Y., Pál, D., Szepesvári, C.: Improved algorithms for linear stochastic bandits. In: NIPS, pp. 2312–2320 (2011)
2. Abe, N., Long, P.M.: Associative reinforcement learning using linear probabilistic concepts. In: ICML, pp. 3–11 (1999)

3. Agarwal, D., Chen, B.C., Elango, P.: Spatio-temporal models for estimating click-through rate. In: WWW, pp. 21–30 (2009)
4. Agrawal, S., Goyal, N.: Thompson sampling for contextual bandits with linear payoffs. In: ICML (3). JMLR Proceedings, vol. 28, pp. 127–135. JMLR.org (2013)
5. Auer, P.: Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research 3, 397–422 (2002)
6. Auer, P., Cesa-bianchi, N., Fischer, P.: Finite-time Analysis of the Multiarmed Bandit Problem. Machine Learning 47, 235–256 (2002)
7. Auer, P., Freund, Y., Schapire, R.E.: The non-stochastic multi-armed bandit problem. Siam Journal on Computing (2002)
8. Chou, K.C., Lin, H.T.: Balancing between estimated reward and uncertainty during news article recommendation for ICML 2012 exploration and exploitation challenge. ICML 2012 Workshop: Exploration and Exploitation 3 (2012)
9. Dani, V., Hayes, T.P., Kakade, S.M.: Stochastic linear optimization under bandit feedback. In: COLT, pp. 355–366 (2008)
10. Li, L., Chu, W., Langford, J., Schapire, R.E.: A contextual-bandit approach to personalized news article recommendation. In: WWW, pp. 661–670 (2010)
11. Cesa-Bianchi, N., Gentile, C., Zappella, G.: A Gang of Bandits. In: NIPS (2013)
12. Wu, F., Huberman, B.A.: Novelty and collective attention. Tech. rep., Proceedings of National Academy of Sciences (2007)
13. Yahoo!: Yahoo! Webscope dataset R6A/R6B. ydata-frontpage-todaymodule-clicks (2011), `http://webscope.sandbox.yahoo.com/`