Enhancing UCB-tuned and Asymptotically Optimal UCB Algorithms through Weighted Average Techniques in Multi-Armed Bandit Scenarios

Chang Qu *

SWJTU-Leeds joint school, Southwest Jiaotong University, Chengdu, 611756, China

* Corresponding Author Email: sc22cq@leeds.ac.uk

Abstract. This paper delves into the complexities of the Multi-Armed Bandit (MAB) problem, a fundamental concept in reinforcement learning and probability theory, with a focus on its application in recommendation systems and dynamic fields such as dynamic pricing and investment. It begins by shedding light on the essential paradox at the heart of the MAB problem – the balance between exploration and exploitation within limited parameters. The study primarily centers on Upper Confidence Bound (UCB) policies, especially UCB-tuned and Asymptotically Optimal UCB, noted for their adept balance between exploration and utilization. The novel contribution of this research is the enhancement of these UCB policies via an innovative weighted average method, leading to the development of WA-UCB-tuned and WA Asymptotically Optimal UCB algorithms. The research rigorously compares these optimized iterations with traditional UCB1, UCB-tuned, and Asymptotically Optimal UCB across varied MAB models featuring different numbers of arms. This study provides an exhaustive introduction to the MAB problem and pertinent UCB policies, the methodology behind the weighted average optimization, extensive experimental analysis, and comprehensive evaluations of the findings. The results showcase marked improvements in algorithmic performance, suggesting significant advancements in the domain of recommendation systems and other applications of the MAB problem.

Keywords: Multi-Armed Bandit Problem; Reinforcement Learning; Upper Confidence Bound; UCB-tuned.

1. Introduction

The Multi-Armed Bandit problem, a cornerstone in reinforcement learning and probability theory, illustrates a critical balance between exploring an environment to identify profitable actions and frequently choosing the empirically optimal action. This problem is extensively applied in recommendation systems, such as on shopping and news websites, and is crucial in dynamic pricing and investment strategies. Numerous algorithms have been developed to maximize performance in this context, including Thompson Sampling, Upper Confidence Bound policies, and the Explore-Then-Commit (ETC) strategy. Particularly, UCB policies are notable for maintaining an effective balance between exploration and exploitation [1].

This study focuses on enhancing two UCB policies: UCB-tuned and Asymptotically Optimal UCB, using a weighted average method. The research introduces the Weighted Average UCB-tuned and Weighted Average Asymptotically Optimal UCB, subsequently comparing these with UCB1, UCB-tuned, and Asymptotically Optimal UCB across various MAB models with differing numbers of arms [2, 3]. The structure of the paper is organized as follows: Initially, foundational theories of the MAB problem and associated UCB policies are presented. Following this, the study integrates a weighted average method to refine these algorithms [4]. The paper then describes the conducted experiments, presenting results and evaluations. The study concludes with a summary of findings and implications.

2. Relevant Theories and Modification

2.1. Multi-Armed Bandit Problem

The Multi-Armed Bandit Problem presents a paradox between exploration and utilization. In a limited number of rounds, learner is supposed to take the optimal choice of a ranged set and maximize the gain (reward) without precognition of the rewards of different actions. The name 'bandit' comes from monitoring a gambler at a row of slot machines in a casino. Gambler's general profit relies on the sequence that the gambler pulls the arms [5]. In MAB problem, 'regret' is defined as the cumulative gap between the reward of each action and the best action in case the best arm is known to the learner. Thus, the ultimate goal for algorithms is also to minimize the cumulative regret value in the whole process.

2.2. UCB, UCB-tuned, and Asymptotically Optimal UCB

To achieve the goal of gaining as much reward as possible, UCB is a classic and efficient algorithm for the MAB problem. Theoretically, UCB polices can guarantee the upper bound of the expected loss, attaching enough importance to exploiting the best action when exploring the sub-optimal arms. If the number of arms is K, then in the first K rounds, the chosen arm A_t follows:

A_t = 1, 2, 3... K. In the remaining rounds, UCB policies take actions (arms) as (1).

$$A_t = \text{argmax}_i \left(\hat{\mu}_i(t-1) + UCB_i(t-1)\right) \tag{1}$$

Where $\hat{\mu}_i(t-1)$ is the mean reward of arm i in these t - 1 rounds, and $UCB_i(t-1)$ is the UCB value of arm i. UCB value is calculated differently in different UCB policies?

UCB1 (Auer et al., 2002) algorithm is the initial version of UCB policies, usually defined as (2).

$$UCB1_i(t-1) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\log n}{T_i(t-1)}} \tag{2}$$

Where t is the serial number of the round, i is the index of the arm, n refers to the total rounds that the algorithm runs in the whole running process, and $T_i(t-1)$ is the number of times that arm i has been chosen until round t - 1.

To make the UCB policies fit the data set collected from the real statistical data, a new parameter B is imported. Then UCB1 is defined as (3).

$$UCB1_i(t-1) = \hat{\mu}_i(t-1) + \frac{B}{2}\sqrt{\frac{2\log n}{T_i(t-1)}} \tag{3}$$

Where B stands for the gap between the possible maximum reward and the possible minimum reward from all of the arms. (The distribution span of rewards).

UCB-tuned is a variation of the UCB1 policy. UCB-tuned adds the variance of the reward into parameters. It is defined as (4) with the parameter B.

$$UCB-tuned_i(t-1) = \hat{\mu}_i(t-1) + \frac{B}{2}\sqrt{\frac{\ln n}{T_i(t-1)}\min\{1/4, V_i(T_i(t-1)\}} \tag{4}$$

Where

$$V_j(T_i(t-1)) \stackrel{\text{def}}{=} \left(\frac{1}{T_i(t-1)}\sum_{\tau=1}^{T_i(t-1)} X_{i,\tau}^2\right) - \hat{\mu}_i{}^2(t-1) + \sqrt{\frac{2\ln n}{T_i(t-1)}} \tag{5}$$

Where $X_{i,\tau}$ refers to the reward from arm i when it is chosen for the τ time.

Asymptotically optimal UCB algorithm concerns with the how algorithm behaves as the number of rounds increases beyond all limits [6].

It is defined as (6) with parameter B.

$$\mathrm{AO}UCB_i(t-1) = \hat{\mu}_i(t-1) + \frac{B}{2}\sqrt{\frac{2log\ (f(t))}{T_i(t-1)}} \tag{6}$$

Where

$$f(t) = 1 + t(\log\ t)^2 \tag{7}$$

2.3. Weighted Method for Algorithm Optimization

In this research, a generalized weighted average method is imported to optimize the parameters. As shown in Fig 1. It is defined as (8) [7].

$$\mu(x, y \mid \alpha, m) = [(1-\alpha)x^m + \alpha y^m]^{1/m} \tag{8}$$

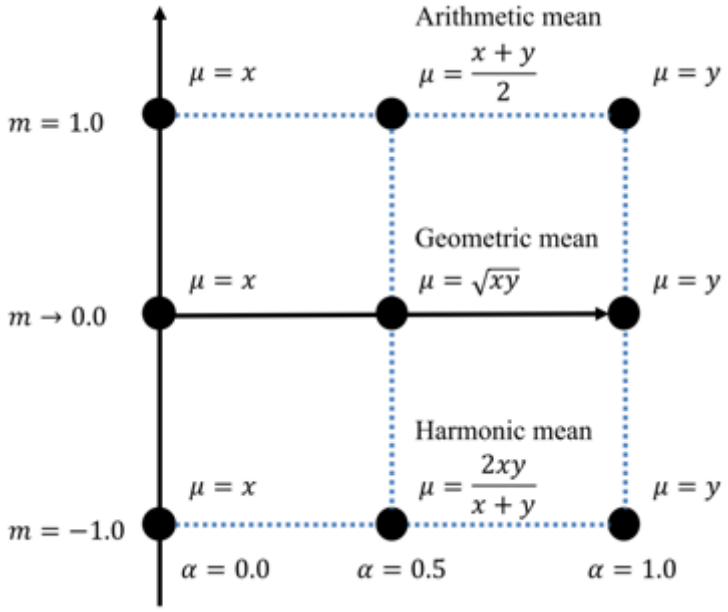

Fig 1. Overview of generalized weighted averages [8].

The application of (8) in the UCB-tuned algorithm and the asymptotically optimal UCB is respectively as (9) and (10).

$$WA - UCB - tuned_i(t-1) = \ (\alpha\ \hat{\mu}_i{}^m(t-1) + (1-\alpha)(\frac{B}{2}\sqrt{\frac{ln\ n}{T_i(t-1)}min\{1/4, V_i(T_i(t-1)\}}\ )^m\ )^{1/m} \tag{9}$$

$$WA - AOUCB_i(t-1) = \ (\alpha\ \hat{\mu}_i{}^m(t-1) + (1-\alpha)(\frac{B}{2}\sqrt{\frac{2log\ (f(t))}{T_i(t-1)}}\ )^m\ )^{1/m} \tag{10}$$

Where $V_i(T_i(t-1)$ in (9) is (5) and $f(t)$ in (10) is (7). When m = 1 and α = 0.5, the WA-UCB algorithms are equivalent to the original UCB algorithms.

3. Experiments

The experiments are based on the data from https://grouplens.org/datasets/movielens/1m/.

The data set has 18 genres which are used as the arms of the bandit where rewards range from 1 to 5. Thus, parameters are set as followed: n = 50000, B = 4 and K = 18 in the absence of special instructions.

3.1. Preliminary Experiment

In order to get the performance of target algorithm in bandits based on this data set in advance and verify the feasibility of the experiment, UCB1, asymptotically optimal UCB , Thompson Sampling and UCB-tuned are compared in the same K = 18 data set where n = 10000. The experiment was performed 100 times and the average of regret values was taken.

In this preliminary experiment, the parameter m and parameter α of the asymptotically optimal UCB and the UCB-tuned are respectively 1 and 0.5. This means they are equivalence to their corresponding original forms.

3.2. Experiment - 1: Optimization of UCB-tuned Using Weighted Average Method

In this experiment, UCB - tuned is optimized by the weighted average method. The parameter α is ranged from 0 to 1 with a span of 0.1, and the parameter m is ranged from 0.2 to 3.8 with a span of 0.3. The experiment is carried out 100 times, and the algorithm ran 50000 rounds each time. Mean cumulative regret values in these 100 times are taken as an indicator of performance evaluation. The parameter combination with the lowest regret value is the optimized WA-UCB-tuned policy.

3.3. Experiment - 2: Optimization of Asymptotically Optimal UCB Using Weighted Average Method

Similar to 3.2 (Experiment - 1), the asymptotically optimal UCB is optimized by the weighted average method. The parameter α is ranged from 0 to 1 with a span of 0.1, and the parameter m is ranged from 0.2 to 3.8 with a span of 0.3. The experiment is also carried out 100 times, and the algorithm ran 50000 rounds each time. The indicator of performance evaluation is the same to that of Experiment – 1 [9]. The parameter combination with the lowest regret value is the optimized WA-Asymptotically optimal UCB (WA-AOUCB) policy.

3.4. Experiment - 3: Behavior of Algorithms in Different Multi-Armed Bandit Environments

In this experiment, the WA-UCB-tuned and WA-AOUCB gotten in the former experiments are tested. This experiment compares the behavior of WA-UCB-tuned, WA-AOUCB, UCB-tuned, AOUCB and UCB1 in a simulation bandit, where K = 2, 8, 32 and 128 [10]. The experiment is also carried out 100 times, and mean regret values of each algorithm are shown in the same graph.

4. Results

The results of 3.1, 3.2, 3.3 and 3.4 are as followed.

The preliminary experiment verifies the the feasibility of the experiment. In this 10000 rounds, 100 times experiment, Asymptotically Optimal UCB shows a larger cumulative regret value than UCB1 and UCB-tuned, and UCB-tuned performs better than UCB1. As shown in Fig 2.

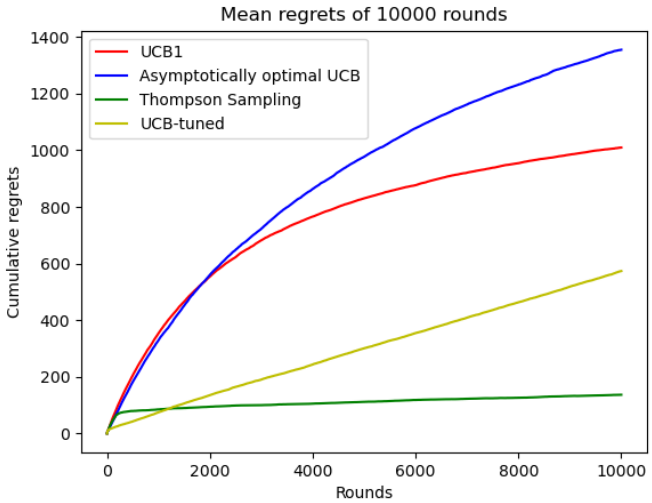

Fig 2. Result of Preliminary Experiment (Photo/Picture credit: Original).

The Experiment - 1 shows the best parameter combination. When α= 0.3 and m = 1.7, the WA-UCB-tuned reaches the lowest regret value. Fig 3 shows the detailed regret values in the experiment - 1. The regret value in the figure is the value after log10 processing.

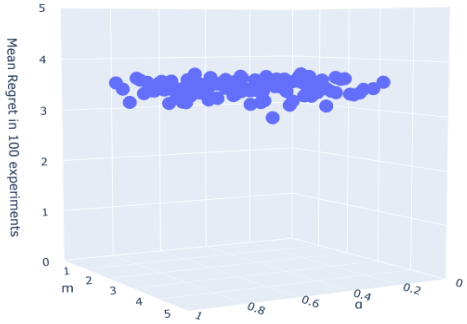

Fig 3. Result of the Experiment - 1 (Photo/Picture credit: Original).

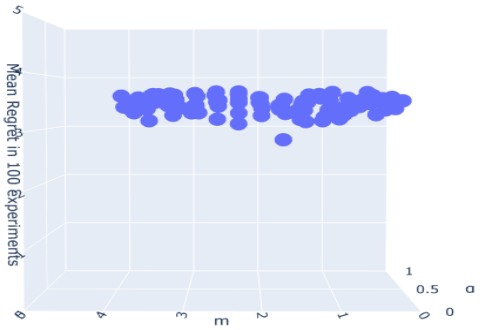

Fig 4. Result of the Experiment - 1 (Photo/Picture credit: Original).

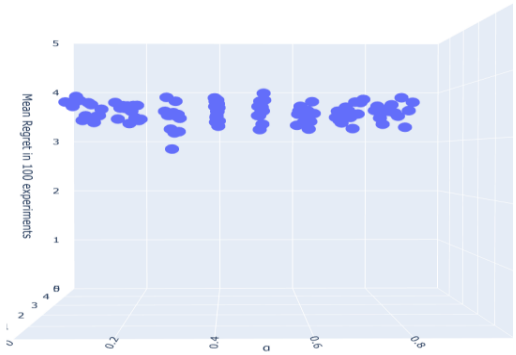

Fig 5. Result of the Experiment - 1 (Photo/Picture credit: Original).

Then the WA-UCB-tuned is defined as (11).

$$WA-UCB-tuned_i(t-1) = \left(\frac{3}{10}\ \hat{\mu}_i^{\frac{17}{10}}(t-1) + \frac{7}{10}\left(\frac{B}{2}\sqrt{\frac{ln\ n}{T_i(t-1)}min\{1/4, V_i(T_i(t-1)\}}\right)^{\frac{17}{10}}\right)^{\frac{10}{17}} \quad (11)$$

As for the Experiment - 2, the WA-AOUCB performs best where α = 0.4 and m = 2. Fig. 4 shows the detailed regret values in the experiment - 2. (The regret value in the figure is the value after log10 processing)

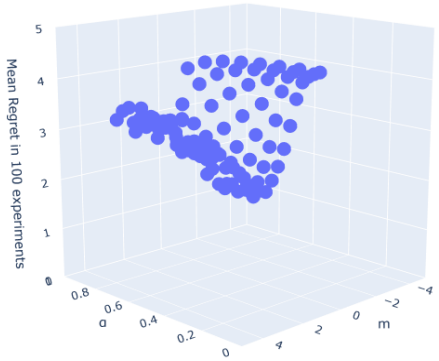

Fig 6. Result of the experiment - 2 (Photo/Picture credit: Original).

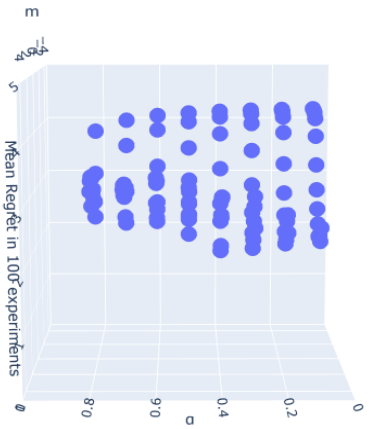

Fig 7. Result of the experiment – 2 (Photo/Picture credit: Original).

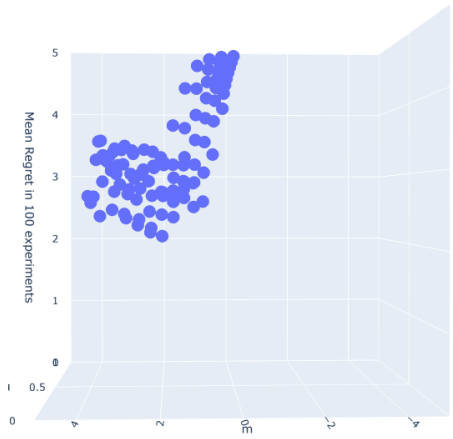

Fig 8. Result of the experiment - 2 (Photo/Picture credit: Original).

And the WA-UCB-tuned is defined as (12).

$$WA - AOUCB_i(t-1) = (\alpha\, \hat{\mu}_i{}^m(t-1) + (1-\alpha)(\frac{B}{2}\sqrt{\frac{2log\,(f(t))}{T_i(t-1)}}\,)^m\,)^{1/m} \quad (12)$$

Where $f(t)$ is defined as (7).

Fig 5 to Fig 8 display the behavior of WA-UCB-tuned, WA-AOUCB, UCB-tuned, AOUCB and UCB1 in different bandits where k = 2, 8, 32, 128 and n = 50000.

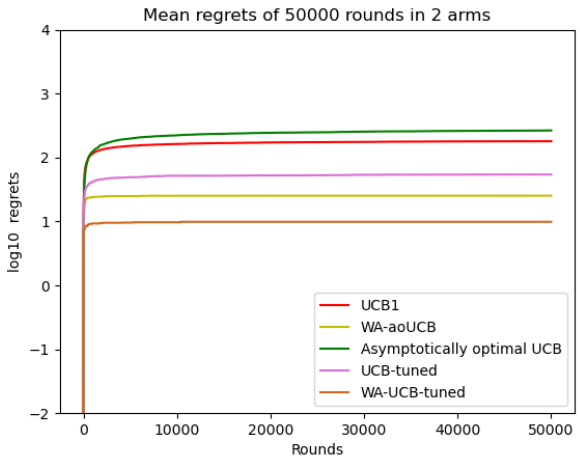

Fig 9. Behavior of UCB policies in 2-armed bandit (Photo/Picture credit: Original).

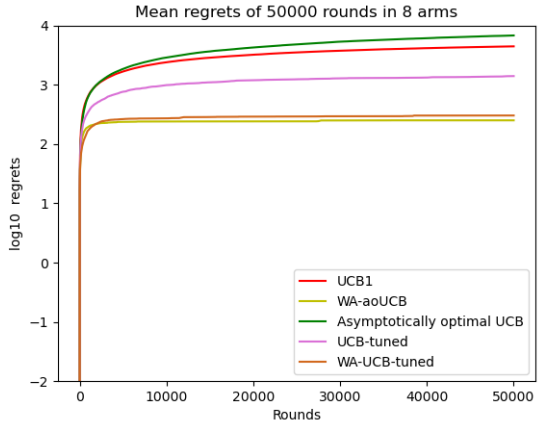

Fig 10. Behavior of UCB policies in 8-armed bandit (Photo/Picture credit: Original).

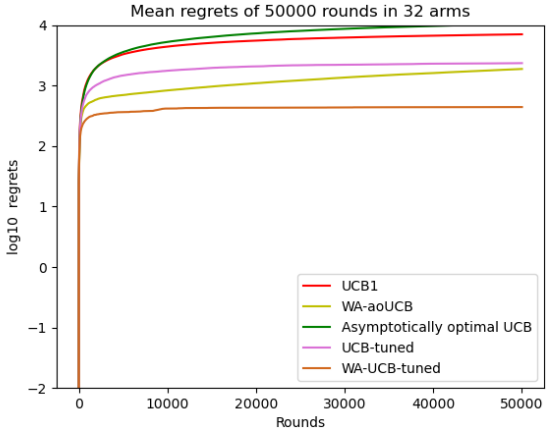

Fig 11. Behavior of UCB policies in 32-armed bandit (Photo/Picture credit: Original).

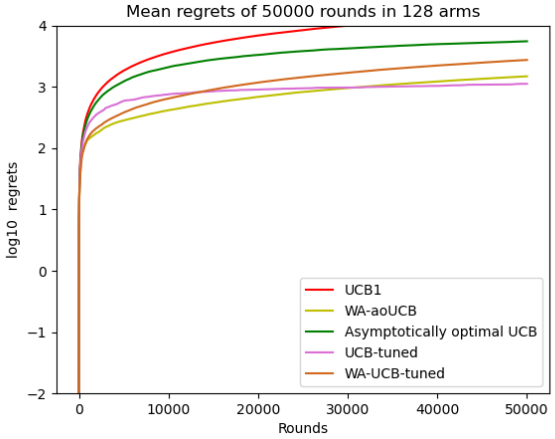

Fig 12. Behavior of UCB policies in 128-armed bandit (Photo/Picture credit: Original).

The results confirm that in comparison to the original forms, both of the WA-UCB-tuned and the WA-AOUCB perform better in bandits with different number of arms. Optimized algorithms efficiently reduce accumulated regret value during the previous exploration. As shown in Fig 9-12.

5. Conclusion

In conclusion, this research has successfully optimized two algorithms for addressing Multi-Armed Bandit (MAB) problems: UCB-tuned and Asymptotically Optimal UCB, utilizing a weighted average method. The study introduces WA-UCB-tuned and WA-AOUCB, corresponding to the enhanced versions of UCB-tuned and Asymptotically Optimal UCB, respectively. Optimal

performance for WA-UCB-tuned is achieved with $\alpha = 0.3$ and $m = 1.7$, while for WA-AOUCB, the best results are observed with $\alpha = 0.4$ and $m = 2$. In simulation tests, both WA-UCB-tuned and WA-AOUCB demonstrate superior performance, evidenced by lower regret values, compared to their original counterparts and the UCB1 algorithm. Looking ahead, future research faces two primary challenges: 1. determining the impact of the parameter B on algorithm performance. 2. Validating the effectiveness of WA-UCB-tuned and WA-AOUCB in Adversarial MAB and Non-stationary Stochastic MAB contexts. Addressing these challenges may necessitate the introduction of new parameters and refinements to algorithmic steps, paving the way for further advancements in this field.

References

[1] Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. Machine learning, 47, 235-256.

[2] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3-4), 285-294.

[3] Kaufmann, E., Cappé, O., & Garivier, A. (2012, March). On Bayesian upper confidence bounds for bandit problems. In Artificial intelligence and statistics (pp. 592-600). PMLR.

[4] Silva, N., Werneck, H., Silva, T., Pereira, A. C., & Rocha, L. (2022). Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. Expert Systems with Applications, 197, 116669.

[5] Manome, N., Shinohara, S., & Chung, U. I. (2023). Simple Modification of the Upper Confidence Bound Algorithm by Generalized Weighted Averages. ar**v preprint ar**v:2308.14350.

[6] Harper, F. M., & Konstan, J. A. The movielens datasets: History and context, Acm transactions on interactive intelligent systems (tiis), 5 (2016). Cited on, 59.

[7] Komiyama, J. (2016). Asymptotically Optimal Multi-armed Bandit Algorithms Aimed at Online.

[8] Amirizadeh, K., & Rajeswari, M. (2015). Accelerated-Greedy Multi Armed Bandit Algorithm for Online Sequential-Selections Applications. J. Softw., 10(3), 239-249.

[9] Gonçalves, R. A., Almeida, C. P., & Pozo, A. (2015). Upper confidence bound (UCB) algorithms for adaptive operator selection in MOEA/D. In Evolutionary Multi-Criterion Optimization: 8th International Conference, EMO 2015, Guimarães, Portugal, March 29--April 1, 2015. Proceedings, Part I 8 (pp. 411-425). Springer International Publishing.

[10] Liu, Y. E., Mandel, T., Brunskill, E., & Popovic, Z. (2014, July). Trading Off Scientific Knowledge and User Learning with Multi-Armed Bandits. In EDM (pp. 161-168).