Augmented Two-Stage Bandit Framework: Practical Approaches for Improved Online Ad Selection

Seowon Han[†], Ryan Lakritz[†] and Hanxiao Wu[†]

Reddit Inc., 303 2nd Street, South Tower, Floor 5, San Francisco, CA 94107

Abstract

In online advertising, maximizing user engagement and advertiser performance hinges on effective ad selection algorithms. Algorithms that tackle Multi-armed bandit problems, such as Thompson Sampling, excel in exploration, but their utilization of contextual information remains limited. Conversely, contextual bandit approaches personalize ad selection by leveraging user and ad-specific features. However, they perform poorly in contexts with limited data and often encounter cold start problems for new ad groups. To address this dilemma, we propose a novel bandit framework that combines context-free and context-aware rewards and is augmented with historical predicted performance, for which we use predicted click-through rate (pCTR) scores. We will refer to this bandit framework as the Augmented Two-Stage Bandit Framework.

Our bandit framework is comprised of two stages. In the first stage, the framework applies context-free Thompson Sampling augmented by historical pCTR scores for initial exploration. The non-contextual bandit algorithm and generalized patterns recognized by our pCTR model allow for effective mitigation of the cold start problem. In the second stage, the framework shifts to a contextual bandit algorithm for refined exploration and exploitation.

We demonstrate the efficacy of our proposed method using extensive simulation and experiments conducted on a real-world ads marketplace at Reddit. Compared to traditional bandit algorithms, our historical pCTR augmented Two-Stage Bandit framework achieves significant improvements in click-through rate. These findings underscore the ability of an Augmented Two-Stage Bandit Framework to enhance online ad selection and improve key performance metrics.

Keywords

Online advertising, bandit algorithms, reinforcement learning, contextual bandits, multi-armed bandits, deep neural networks, ad optimization, ad selection, ad retrieval

1. Introduction

This paper introduces a novel Augmented Two-Stage Bandit Framework that improves click-through-rate prediction. Our framework aims to optimize ad performance while handling the inherent data sparsity that is introduced by newly formed ads. Our framework brings two major novelties: a two-stage approach and pCTR (predicted click-through rate) augmentation. The former helps performance in data-sparse environments while the latter further helps with data-sparsity and improves overall outcomes.

In the initial stage, we leverage a context-free bandit algorithm enhanced by historical pCTR scores to effectively explore candidate ads, even with limited data. The context-free bandit algorithm we use is Thompson Sampling, but the general framework is adaptable to any context-free bandit algorithm. This mitigates the cold start problem for new ads, allowing them to quickly learn and adapt to user preferences.

As data accumulates, the framework transitions to a contextual bandit algorithm. We use Linear Thompson Sampling to incorporate user and ad-specific features for refined exploration and exploitation. This ensures personalized ad selection that maximizes click-through rates and improves overall campaign performance.

Our proposed two-stage approach effectively addresses data sparsity and imbalance while reaping the benefits of personalized, context-aware selection with prior knowledge about predicted performance. We demonstrate the efficacy of this method through extensive simulations and real-world experiments on a large-scale ads marketplace. Compared to traditional contextual bandit algorithms, the proposed Augmented Two-Stage Bandit Framework achieves significant improvements in click-through rate and other key performance metrics, underscoring its potential to improve online ad selection.

We delve into the challenges of data sparsity and cold starts, showcase the benefits of our pCTR-augmented approach, and present the compelling results of our experiments. By bridging the gap between exploration and exploitation, while enabling personalization and context-awareness, the proposed method solves a significant issue in intelligent ad selection today, ultimately benefiting both users and advertisers in the online advertising landscape.

2. Related Work

Extensive research exists on exploration versus exploitation algorithms.

Multi-Armed Bandit Algorithms (MABs): Many solutions to the online ad selection problem fall squarely within the domain of bandit algorithms. Traditional algorithms that address the Multi-Armed Bandit problem, like Thompson Sampling, offer efficient exploration-exploitation trade-offs for selecting ads in dynamic environments. However, the lack of context-awareness limits their ability to personalize recommendations based on user and ad-specific features [1, 2].

Contextual Bandit Algorithms: Recognizing the limitations of MABs, contextual bandit solutions, such as LinUCB and Linear Thompson Sampling, are able to leverage additional information, such as user demographics and ad context, for personalized selection[1]. These features are typically integrated through embedding techniques or neural networks, such as deep neural networks (DNNs) [3], which enable the model to learn complex interactions between the contextual information and ad performance [4]. While demonstrating superior performance compared to MABs, their reliance on sufficient per-context data can hinder their effectiveness in data-scarce scenarios [5], which

AdKDD'24: Barcelona, Spain

[†]These authors contributed equally.

samantha.han@reddit.com (S. Han); ryan.lakritz@reddit.com (R. Lakritz); sylvia.wu@reddit.com (H. Wu)

introduces cold start issues in ad selection.

Cold Start Solutions: Various approaches have been explored to address the cold start problem in bandit problem formulations[6, 7, 8, 2, 5]. Bayesian approaches leverage prior information from similar contexts to inform initial decision-making. Others utilize Thompson Sampling with confidence bounds to prioritize exploration for new items. However, these methods often rely on strong assumptions about data similarity or require careful parameter tuning, limiting their ability to generalize [5].

Deep Learning Integration: Recently, the integration of deep neural networks (DNNs) with bandit algorithms has shown promising results for improving ad selection performance [9, 10]. Proposed frameworks integrate these features in a non-linear manner, enabling principled exploration through techniques like neural collaborative filtering recommendation and inference time dropout, leading to improved performance in various applications [11, 12]. However, their computational complexity and potential for overfitting remain challenges to be addressed.

Our proposed Augmented Two-Stage Bandit Framework combines the strengths of context-free bandits for exploration and historical pCTR scores for cold start mitigation, and we aim to:

- **Bridge the gap between exploration and exploitation**: Our approach effectively explores the ad space in the initial phase with minimal data, while seamlessly transitioning to context-aware decision-making as data accumulates [13].
- **Address the cold start problem**: The two-stage approach with pCTR augmentation mitigates the data sparsity issue for new agents, enabling them to quickly learn and adapt to user preferences.
- **Improve overall ad performance**: By personalizing recommendations based on user and ad features, our framework seeks to achieve significant improvements in click-through rates and other key performance metrics compared to existing approaches. We selected features with a total cardinality of less than 10^1 across all features, based on offline feature importance analysis and online experimentation.

The proposed bandit framework also has some drawbacks to consider. It does not solve the problem that bandit algorithms, in general, do not scale well with a large number of features. Another thing to consider is that it's more difficult to do an unbiased offline evaluation of bandit performance, compared to prediction models. This might lead to more work to develop a trustworthy offline evaluation framework or run more online experimentation, which comes with a cost.

3. Problem

3.1. Background on Ad Auction Funnel

We apply an ad selection model as part of our ad auction funnel. Its primary function is to select a subset of candidate ads that will proceed to lower-funnel steps, including inference by a heavy ranking (pCTR) model and final auction ranking. The ad selection stage is necessary due to infrastructure constraints and the amount of computation resources necessary to complete the lower-funnel steps. It is not cost-efficient or technically feasible to apply the pCTR model to all candidate ads. As such, the ad selection model is an important component of our ad auction framework and requires a unique modeling solution.

3.2. Problem Framing

Ad selection can be formulated as a contextual bandit problem. At time t, an ad impression is requested by a user. The decision-making agent of the bandit will observe the feature vector of the impression, which is considered as context $s_t \in S$, and be presented with a set of eligible ads to choose from action space A. The chosen ad, a_t, is the action taken by the agent, where $a_t \in A$. The agent uses the learned selection policy π to choose an ad $a_t = \pi(s_t)$ for the ad impression at time t to maximize the expected click. The system collects the clicks generated by the agent's chosen ad, denoted as reward $r(s, a)$. The given policy's expected reward is denoted as $r_\pi(s, a) = E[r(s, a)]$. For each observation at time t agent improves its selection policy based on the observation tuple $< s_t, a_t, r_\pi(s_t, a_t) >$.

The objective is to find an optimal policy π that maximizes the expected total reward $E[\sum_{t=1}^{T} r(s_t, a_t)]$. During ad selection, the agent will select an ad that maximizes the expected reward given the observed context of an impression: $a_t := \arg\max_{a_t \in A} E[\sum_{t=1}^{T} r(s_t, a_t)]$. The subsequent sections will delve into the methodologies employed to enhance the policy, ultimately aiming for improved click performance.

4. Methods

4.1. Two-Stage Bandit Framework

One notable constraint of contextual bandit algorithms is high variance at the initial stages of learning and, as a result, can over-emphasize exploration. Excessive exploration is costly in an ads marketplace, resulting in sub-optimal ad performance.

To tackle this challenge, we introduce a two-stage bandit framework that optimizes $r_{TS}(s, a)$. The framework consists of a context-free reward $r_{MAB}(a)$ at the initial stage, and a context-aware reward $r_{CB}(s, a)$, which it switches to as data is accumulated. The motivation behind this approach is that in ad selection, the best performing ad for the overall marketplace is likely a better-than-average candidate under different contexts. By initially relying on the context-free policy's rewards when the context information is sparse, and then transitioning to the context-aware policy's rewards once it outperforms the context-free bandit policy, our proposed method aims to improve the performance at the early stage while preserving the advantages of personalized recommendations in the long run.

In order to switch between the context-free and contextual reward, the variance is introduced to the framework as a measure of uncertainty and a degree of exploration. We utilize variance of the expected contextual rewards collected from the first observation through time $t-1$: $Var_t(s, a) = Var_{1 \le i \le t-1, (s_i, a_i) = (s, a)}(r_{CB}(s_i, a_i))$. This yields the variance of the contextual rewards for the given state-action pair up to time $t - 1$.

The $r_{TS}(s, a)$ uses context-free reward until $Var_t(s, a))$ is below a threshold, denoted as τ. In practice, the threshold τ was tuned using offline evaluation and online experimentation, such that it maximized the total reward of the Two-

Stage framework. We started with a wide range of testable τ values and narrowed down to values that indicated stability. From that point, we conducted online experiments which tested smaller adjustments in τ to find an optimally tuned value.

$$r_{TS}(s_t, a_t) = \begin{cases} r_{MAB}(a_t) & \text{if } Var_t(s,a) > \tau \\ r_{CB}(s_t, a_t) & \text{otherwise} \end{cases} \quad (1)$$

4.2. Historical Predicted Click Performance

Our ad auction pipeline relies on the pCTR model to predict click-through rates, which has a deep neural network (DNN) model architecture. This architecture allows us to capture complex interactions between user and ad features and adapt to changing user behavior to enhance the accuracy of click-through rate predictions.

In the heavy ranking stage, the pCTR model is used because it can achieve higher accuracy with sufficient data and features. It is more efficient at incorporating contexts, such as user-specific and interaction features. However, a complex model like pCTR has its drawbacks, including latency and cost, as it requires more time and computational resources to train, maintain, and use for inference. Inference with the pCTR model in real-time for all ads entering the auction pipeline is computationally infeasible, given the latency constraints.

In the ad selection stage of the early auction pipeline, the pCTR models — as well as other non-bandit algorithms — can suffer from feedback loops and selection bias [14]. Bandit algorithms are well-suited for ad selection in this early stage, where exploration is crucial, due to their emphasis on exploration versus exploitation. Additionally, bandit algorithms are advantageous in ad marketplaces where ads can be created or changed at any time, due to their real-time adaptability. These reasons are what lead us to utilizing a bandit algorithm, while incorporating some key strengths of the pCTR model.

4.3. Incorporating Two-Stage Approach and pCTR into our Framework

To address this limitation, we explored an alternative approach that augments our two-stage framework with the pCTR model. We observed a remarkably high correlation (r-squared = 0.9907) between pCTR and the estimated CTR of the following day, suggesting that the pCTR scores from the previous day can still effectively capture the underlying patterns and trends in user behavior. By incorporating pCTR scores as weights in our ad selection bandit framework, we harness the strengths of both models to improve the accuracy and efficiency of our ad auction pipeline.

We introduce the previous day's, pre-computed pCTR scores as weights to the policy's reward with a multiplicative application for the context-free stage. Once the variance of the contextual bandit agent crosses the threshold τ, the framework switches to the context-aware stage. pCTR augmentation is no longer used at this stage in order to reduce infrastructure cost and improve latency.

The final reward function of our pCTR-augmented two-stage bandit framework, denoted $r_{A-TS}(s,a)$ is defined as:

$$r_{A-MAB}(a_t) = r_{MAB}(a_t) * pCTR(a_t) \quad (2)$$

$$r_{A-TS}(s_t, a_t) = \begin{cases} r_{A-MAB}(a_t) & \text{if } Var_t(s,a) > \tau \\ r_{CB}(s_t, a_t) & \text{otherwise} \end{cases} \quad (3)$$

5. Model Training and Serving

5.1. Training

Establishing an online Reinforcement Learning environment for linear bandit algorithms is a significant infrastructure investment that is difficult to balance with serving latency constraints at scale. To mitigate these risks, we opted for a mini-batch training approach, with a training intervals of one hour. This training interval was decided based on an offline simulation.

The offline simulation compares the performance of agents that are retrained at different frequencies. Figure 1 below shows the normalized average reward, i.e. the simulated reward of the given agents divided by the simulated reward of an agent that is retrained at real-time. One hour provided the best performance-to-cost balance, showing marginal drop-off from 15 minutes but a large improvement over daily retraining.

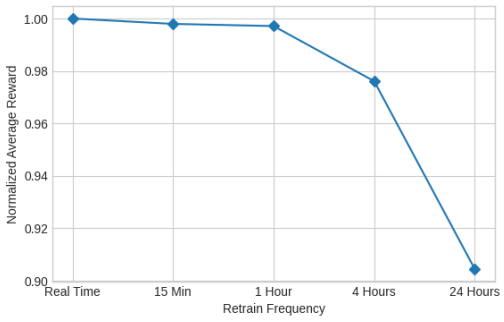

Figure 1: Average Reward at Given Retrain Frequencies

5.2. Serving

As noted, the auction pipeline has very tight serving latency constraints in an online ad serving environment. Our model inference service, which is called at every impression, contains a framework that gets features from our online feature store, routes requests between different models to compute model inferences from our Augmented Two-Stage Bandit Framework.

With our framework, we are able to achieve sub-2*ms* model inference time while serving in production.

6. Experimentation

6.1. Methodology and Dataset

In order to evaluate the impact of our proposed method in a real-world setting, we conduct an online experiment. Our control variant is the Contextual Bandit algorithm without alterations. We have three treatment variants: Two-Stage Bandit, Augmented Contextual Bandit, and Augmented Two-Stage Bandit. All of the models are initialized from a cold-

start and are updated at an hourly cadence. The details of each model are illustrated in the next section.

We use data collected in real-time from Reddit's ad impressions. To handle business logic within the auction system, the model is designed to evaluate at the ad group level; the model chooses an ad with the highest expected reward within an ad group for all ad groups passing through this part of the ad funnel. As such, we only include ad groups with more than one ad for the analysis. Then we evaluate the lift in click-through rate (CTR) of proposed methods versus control variant.

The A/B experiment occurred over 7 days to achieve statistical power and significance on the key metrics and to account for weekly seasonality.

6.2. Model Variants

The following models are included in the online experiment.

Contextual Bandit: In the control variant, a contextual bandit algorithm, specifically, Linear Thompson Sampling [15], is used to select one ad per ad group for the ad selection requests. The control model has the same contextual awareness features and click-based reward compared to the following treatment variants, but it does not apply two-stage framework or augmentation.

Two-Stage Bandit: This is the bandit framework introduced in Section 4.1 Equation 4.1. It relies on Thompson Sampling during the initial stage and transitions to Linear Thompson Sampling as the variance of the contextual reward is below the threshold τ.

Augmented Contextual Bandit: This is the Linear Thompson Sampling algorithm, with rewards that are augmented by pCTR.

$$r_{A-CB}(s,a) = r_{CB}(s,a) * pCTR(a) \tag{4}$$

Augmented Two-Stage Bandit: This is the proposed bandit framework introduced in Section 4.3 Equation 3. This framework applies the pCTR augmentation to context-free Thompson Sampling and switches to Linear Thompson Sampling when the variance of the contextual reward is below the threshold τ.

7. Results

7.1. Aggregate Results

Table 1 summarizes the lift in click-through rate, a key performance metric in ads marketplace, and compares three test variants against the control model. The results indicate a marginal decline in lift for the Augmented Contextual Bandit, whereas the Augmented Two-Stage Bandit exhibits a modest increase in click-through rate (CTR).

	CTR Lift
Control	-
Augmented Two-Stage Bandit	0.97%
Two-Stage Bandit	0.49%
Augmented Contextual Bandit	-0.12%

Table 1
Relative CTR lift versus Control

Notably, our proposed method, which integrates the two-stage bandit framework and pCTR augmentation, achieves the highest overall performance across all ad groups. Our bandit framework effectively mitigates the cold-start problem, enabling better CTR performance even during early exploration. In contrast, the Augmented Contextual Bandit exhibited slight negative lift, highlighting the significance of integrating the two-stage bandit framework in achieving these gains. Ultimately, the combination of historical pCTR augmentation and the two-stage bandit framework demonstrates a combined effect that surpasses the performance of either approach in isolation.

7.2. CTR Lift in Data-Scarce Ad Groups

Figure 2 provides important insight into performance in cold start and data-scarcity by segmenting the performance using impressions per ad within the ad group. The ad groups with low number of impressions per ad indicate that the algorithms have less data to train and are prone from heavy exploration in contextual bandit algorithms. In Figure 2, each dot presents the lift for ad groups within the percentile of cumulative impressions per ad within the ad group.

The Augmented Two-Stage Bandit performs significantly better than baseline, particularly for low impression ad groups. Specifically, for the "p10 Ad Groups" - which comprise of ad groups with impressions per ad below the 10^{th} percentile - our approach achieves 10.95% improvement in CTR compared to the baseline.

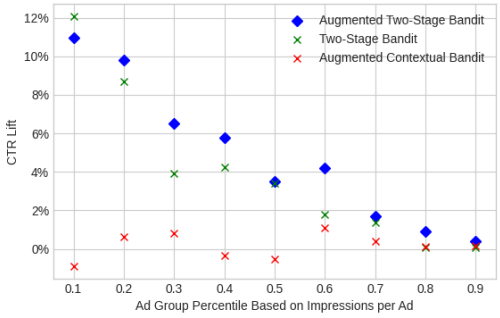

Figure 2: CTR Lift at Ad Group Percentile Based on Impressions per Ad

7.3. Cumulative CTR Lift Over Time

Figure 3 shows the cumulative daily lift achieved by our proposed methods. Notably, the Augmented Two-Stage Bandit exhibits a pronounced lift during the initial two days of the experiment where most of the ad groups have low impression counts. This highlights the improved cold-start performance at the onset of model deployment. As more data is accumulated, the Augmented Two-Stage Bandit Framework switches to Contextual Bandit and the lift decreases as expected.

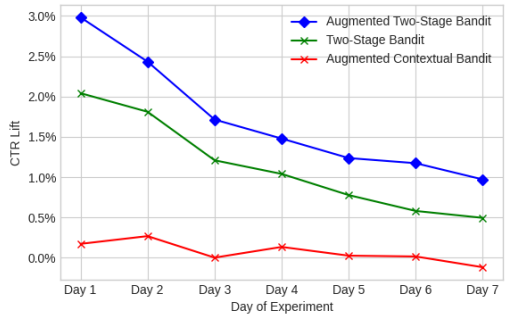

Figure 3: Cumulative CTR Lift versus Control

7.4. Cumulative Click Volume Lift Over Time

In addition to the lift in click-through rate, Figure 4 presents a comparative analysis of click volume across the different variants, further reinforcing the benefits of the proposed approach. Our proposed strategy has consistently generated the highest click volume throughout the experiment. The increase in click volume, coupled with the lift in CTR underscores the holistic performance improvement achieved by the proposed solution.

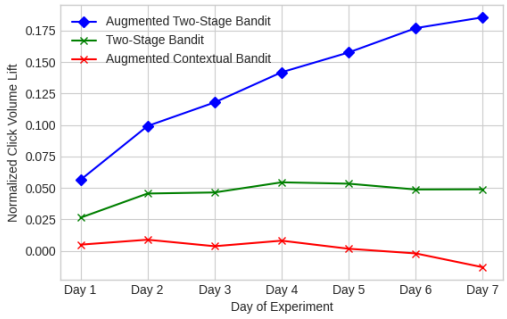

Figure 4: Cumulative Lift in Click Volume versus Control

8. Conclusion

In this paper, we define a framework for improving upon linear contextual bandit algorithms for online ad selection, particularly by focusing on performance in cold-start and data-scarce scenarios. The Augmented Two-Stage Bandit Framework is a novel approach to selecting personalized ads while leveraging exploration to address the cold-start problem present in many personalized recommendation models. Our framework showed a significant CTR lift in experiments, with especially large improvements in ad groups with fewer impressions. Our framework offers practical application to online serving with low-latency requirements significantly improving key performance metrics in our ad marketplace.

9. Acknowledgements

Thank you to Bee Massi, Simon Kim, and Josh Cherry for reviewing and advising on the underlying two-stage and augmentation approaches that are utilized in this paper.

References

[1] J. Langford, T. Zhang, The epoch-greedy algorithm for multi-armed bandits with side information, in: J. Platt, D. Koller, Y. Singer, S. Roweis (Eds.), Advances in Neural Information Processing Systems, volume 20, Curran Associates, Inc., 2007. URL: https://proceedings.neurips.cc/paper_files/paper/2007/file/4b04a686b0ad13dce35fa99fa4161c65-Paper.pdf.

[2] S. Caron, S. Bhagat, Mixing bandits: a recipe for improved cold-start recommendations in a social network, in: Proceedings of the 7th Workshop on Social Network Mining and Analysis, SNAKDD '13, Association for Computing Machinery, New York, NY, USA, 2013. URL: https://doi.org/10.1145/2501025.2501029. doi:10.1145/2501025.2501029.

[3] J. Chen, B. Sun, H. Li, H. Lu, X.-S. Hua, Deep ctr prediction in display advertising, in: Proceedings of the 24th ACM International Conference on Multimedia, MM '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 811–820. URL: https://doi.org/10.1145/2964284.2964325. doi:10.1145/2964284.2964325.

[4] D. Chakrabarti, D. Agarwal, V. Josifovski, Contextual advertising by combining relevance with click feedback, in: Proceedings of the 17th International Conference on World Wide Web, WWW '08, Association for Computing Machinery, New York, NY, USA, 2008, p. 417–426. URL: https://doi.org/10.1145/1367497.1367554. doi:10.1145/1367497.1367554.

[5] N. Silva, T. Silva, H. Werneck, L. Rocha, A. Pereira, User cold-start problem in multi-armed bandits: When the first recommendations guide the user's experience, ACM Trans. Recomm. Syst. 1 (2023). URL: https://doi.org/10.1145/3554819. doi:10.1145/3554819.

[6] L. Guo, J. Jin, H. Zhang, Z. Zheng, Z. Yang, Z. Xing, F. Pan, L. Niu, F. Wu, H. Xu, C. Yu, Y. Jiang, X. Zhu, We know what you want: An advertising strategy recommender system for online advertising, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 2919–2927. URL: https://doi.org/10.1145/3447548.3467175. doi:10.1145/3447548.3467175.

[7] F. Pan, S. Li, X. Ao, P. Tang, Q. He, Warm up cold-start advertisements: Improving ctr predictions via learning to learn id embeddings, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 695–704. URL: https://doi.org/10.1145/3331184.3331268. doi:10.1145/3331184.3331268.

[8] H. T. Nguyen, J. Mary, P. Preux, Cold-start problems in recommendation systems via contextual-bandit algorithms, CoRR abs/1405.7544 (2014). URL: http://arxiv.org/abs/1405.7544. arXiv:1405.7544.

[9] D. Zhou, L. Li, Q. Gu, Neural contextual bandits with UCB-based exploration, in: H. D. III, A. Singh

(Eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 11492–11502. URL: https://proceedings.mlr.press/v119/zhou20a.html.

[10] Q. Shi, F. Xiao, D. Pickard, I. Chen, L. Chen, Deep neural network with linucb: A contextual bandit approach for personalized recommendation, in: Companion Proceedings of the ACM Web Conference 2023, WWW '23 Companion, Association for Computing Machinery, New York, NY, USA, 2023, p. 778–782. URL: https://doi.org/10.1145/3543873.3587684. doi:10.1145/3543873.3587684.

[11] M. Collier, H. U. Llorens, Deep contextual multi-armed bandits, CoRR abs/1807.09809 (2018). URL: http://arxiv.org/abs/1807.09809. arXiv:1807.09809.

[12] M. Unger, A. Tuzhilin, A. Livne, Context-aware recommendations based on deep learning frameworks, ACM Trans. Manage. Inf. Syst. 11 (2020). URL: https://doi.org/10.1145/3386243. doi:10.1145/3386243.

[13] S. Groman, B. Massi, S. Mathias, D. Lee, J. Taylor, Model-free and model-based influences in addiction-related behaviors, Biological Psychiatry 85 (2019). doi:10.1016/j.biopsych.2018.12.017.

[14] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, X. He, Bias and debias in recommender system: A survey and future directions, CoRR abs/2010.03240 (2020). URL: https://arxiv.org/abs/2010.03240. arXiv:2010.03240.

[15] S. Agrawal, N. Goyal, Thompson sampling for contextual bandits with linear payoffs, CoRR abs/1209.3352 (2012). URL: http://arxiv.org/abs/1209.3352. arXiv:1209.3352.