Designing Truthful Contextual Multi-Armed Bandits based Sponsored Search Auctions

Extended Abstract

Kumar Abhishek
International Institute of Information Technology (IIIT)
Hyderabad, India
kumar.abhishek@research.iiit.ac.in

Shweta Jain
Indian Institute of Technology (IIT)
Ropar, India
shwetajain@iitrpr.ac.in

Sujit Gujar
International Institute of Information Technology (IIIT)
Hyderabad, India
sujit.gujar@iiit.ac.in

ABSTRACT

We consider the Contextual Multi-Armed Bandit (ConMAB) problem for sponsored search auction (SSA) in the presence of strategic agents. The problem has two main dimensions: i) Need to learn unknown click-through rates (CTR) for each agent and context combination and ii) Elicit true bids from the agents. Thus, we address the problem to design non-exploration-separated truthful MAB mechanism in the presence of contexts (aka side information). Towards this, we first design an elimination-based ex-post monotone algorithm *ELinUCB-SB*, thus leading to an ex-post incentive compatible mechanism. *M-ELinUCB-SB* outperforms the existing mechanisms available in the literature; however, theoretically, the mechanism may incur linear regret in some instances. We next design *SupLinUCB*-based allocation rule *SupLinUCB-S* which obtains a worst-case regret of $O(n^2\sqrt{dT \log T})$ as against $O(n\sqrt{dT \log T})$ for non-strategic settings; $O(n)$ is price of truthfulness. We demonstrate the efficacy of both of our mechanisms via simulation and establish superior performance over the existing literature.

KEYWORDS

Contextual Multi-armed Bandits, Mechanism Design

ACM Reference Format:
Kumar Abhishek, Shweta Jain, and Sujit Gujar. 2020. Designing Truthful Contextual Multi-Armed Bandits based Sponsored Search Auctions. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

The probability of an ad gets clicked, referred to as *click-through rate* (CTR), plays a crucial role in SSA. The CTR of an ad is unknown to the center (auctioneer), but it can learn CTRs by displaying the ad repeatedly over a period of time. Each agent i also has a private valuation of v_i for its ad, which represents its willingness to pay for a click. This valuation needs to be elicited from the agents truthfully.

In the absence of contexts, if the agents report their real valuations, we can model the problem as a Multi-Armed Bandit (MAB) problem [9] with agents as arms. To elicit truthful bids from the agents, we can use *Mechanism Design* [2, 11]. Such mechanisms are oblivious to the learning requirements and fail to avoid manipulations by the agents when learning is involved. In such cases, the researchers have modeled this problem as a *MAB mechanism* [4–8, 10, 12]. The authors designed *ex-post truthful (–incentive-compatible)* (EPIC) mechanisms wherein the agents are not able to manipulate even when the random clicks are known to them. To the best of our knowledge, contextual information in SSA is considered only in [6]. The authors proposed a deterministic, exploration-separated mechanism (we call it *M-Reg*) that offers strong game-theoretic properties. However, it faces multiple practical challenges like high regret, prior knowledge of the number of rounds, and exploration-separateness, which can cause agents to drop off after some rounds. We resolve in this paper in the next section.

2 MODEL AND ALGORITHMS

Consider a fixed set of agents $\mathcal{N} = \{1, 2, \dots, n\}$, with each agent having exactly one ad competing for a single slot available to the center. Before the start of the auction, each agent i submits the valuation of getting a click on its ad as bid b_i. A contextual n−armed MAB mechanism $\mathcal{M}$ proceeds in discrete rounds $t = 1, 2, \dots, T$. At each round t:

(1) $\mathcal{M}$ observes a context $x_t \in [0, 1]^d$ which summarizes the profile of the user arriving at round t.
(2) Based on the history, h_t, of allocations, observed clicks, and the context x_t, $\mathcal{M}$ chooses an agent $I_t \in \mathcal{N}$.
(3) $\mathcal{M}$ observes r_{I_t} which is 1 if it gets clicked and 0 otherwise. No feedback on the other agents.
(4) $\mathcal{M}$ determines payment $p_{I_t,t} \geq 0$ that I_t pays to the center. The payments of other agents are 0.
(5) Update $h_t = h_{t-1} \cup \{x_t, \{I_t\}, \{r_{I_t}\}\}$.
(6) $\mathcal{M}$ improves arm-selection strategy with new observation.

To capture contextual information, we assume that the CTR of an agent i is linear in d-dimensional context x_t with some unknown coefficient vector θ_i. Thus CTR for agent i at given round t is: $\mu_i(x_t) = \mathbb{P}[r_{i,t}|x_t] = \theta_i^{\mathsf{T}} x_t$. The objective of $\mathcal{M}$ is to minimize social welfare *regret* which is given as:

$$\mathbb{R}_T(\mathcal{M}) = \sum_{t=1}^{T} [\theta_{i_t^*}^T x_t \cdot b_{i_t^*} - \theta_{I_t}^T x_t \cdot b_{I_t}] \qquad (1)$$

Here, $i_t^*(x_t) = argmax_k\{b_k \cdot (\theta_k^T x_t)\}$.

We next present our algorithms, namely *ELinUCB-SB* and *SupLinUCB-S* satisfying *ex-post monotonicity*, i.e., each agent's number of clicks increases with the increase in bid irrespective of the contextual information and random realization of clicks.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

Algorithm 1 *ELinUCB-SB*

1: **Inputs:** $n, T, \alpha \in \mathbb{R}_{+}$, bid vector b, batch size bs
2: **Initialization:** $S_{act} = \mathcal{N}, x^{'} \leftarrow 0_{d\times 1}, T^{'} = \lfloor \frac{T}{bs} \rfloor$
3: **for all** $i \in \mathcal{N}$ **do**
4: $A_i \leftarrow I_d$ (d-dimensional identity matrix)
5: $c_i \leftarrow 0_{d\times 1}$ (d-dimensional zero vector)
6: $\mu_i^{+} \leftarrow b_i; \mu_i^{-} \leftarrow 0$
7: **for** $t^{'} = 1, 2, 3, \ldots, T^{'}$ **do**
8: $I_{t^{'}} \leftarrow 1 + (t^{'} - 1) \mod n$
9: **if** $I_{t^{'}} \in S_{act}$ **then**
10: **for** $t = (t^{'} - 1)bs, \ldots, (t^{'} \cdot bs - 1)$ **do**
11: Observe context as x_t
12: $I_t \leftarrow I_{t^{'}}$,
13: $x^{'} \leftarrow ((t-1)x^{'} + x_t)/t$ (averaging over contexts)
14: Observe click as $r_{I_t} \in \{0, 1\}$
15: $A_{I_t} \leftarrow A_{I_t} + x_t x_t^{\intercal}, c_{I_t} \leftarrow c_{I_t} + r_{I_t} x_t, \hat{\theta}_{I_t} \leftarrow A_{I_t}^{-1} c_{I_t}$
16: **if** $\mu_{I_t}^{-} < \mu_{I_t}^{+}$ **then**
17: $(\gamma_{I_t}^{-}, \gamma_{I_t}^{+}) \leftarrow b_{I_t}(\hat{\theta}_{I_t}^{\intercal} x^{'} \mp \alpha \sqrt{(x^{'})^{\intercal} A_{I_t}^{-1} x^{'}})$
18: **if** $\max(\mu_{I_t}^{-}, \gamma_{I_t}^{-}) < \min(\mu_{I_t}^{+}, \gamma_{I_t}^{+})$ **then**
19: $(\mu_{I_t}^{-}, \mu_{I_t}^{+}) \leftarrow (\max(\mu_{I_t}^{-}, \gamma_{I_t}^{-}), \min(\mu_{I_t}^{+}, \gamma_{I_t}^{+}))$
20: **else**
21: $\left(\mu_{I_t}^{-}, \mu_{I_t}^{+}\right) \leftarrow \left(\frac{\mu_{I_t}^{-} + \mu_{I_t}^{+}}{2}, \frac{\mu_{I_t}^{-} + \mu_{I_t}^{+}}{2}\right)$
22: **else**
23: **for** $t = (t^{'} - 1)bs, \ldots, (t^{'} \cdot bs - 1)$ **do**
24: Observe x_t
25: $I_t \leftarrow argmax_i\, b_i \cdot (\hat{\theta}_i^T x_t), \ \ni I_t \in S_{act}$
26: Observe click as $r_{I_t} \in \{0, 1\}$
27: **for all** agent $i \in S_{act}$ **do**
28: **if** $\mu_i^{+} < \max_{k \in S_{act}} \mu_k^{-}$ **then**
29: Remove i from S_{act}

Intuitiion behind *ELinUCB-SB*: The algorithm maintains a set of active agents S_{act}. Once an agent is evicted from S_{act}, it can not be added back. At each round t, the algorithm observes context x_t. It determines the index of agent $I_{t'}$ whose turn is to display the ad based on round robin order (line[8]). The algorithm then checks if $I_{t'} \in S_{act}$. If it evaluates to true the algorithm does exploration (lines[9-21]) else exploitation (lines[23-26]). It is important to note that no parameter is updated during exploitation, which is crucial for the ex-post monotonicity property. At the end of each round, elimination (lines[27-29]) is done which removes the agents $j \in S_{act}$ from S_{act} if UCB of agent j is less than LCB of any other agent in S_{act}. Update on bounds over the average of context after the completion of batch allocation handles the variance in contexts and its arrivals, thus reducing the regret significantly. It can be shown that eventually, *ELinUCB-SB* will eliminate all but one arm. Even though *ELinUCB-SB* incurs linear regret theoretically, it performs well in simulation and has interesting monotonicity properties. Similarly, *SupLinUCB-S* is derived from *SupLinUCB* to ensure ex-post monotonicity.

Theorem 2.1. *The allocation rules induced by ELinUCB-SB (Algorithm 1) and SupLinUCB-S (Algorithm 2) are ex-post monotone.*

Theorem 2.2. *SupLinUCB-S has regret* $O(n^2\sqrt{dT \ln T})$ *with probability at least* $1 - \kappa$ *if it is run with* $\alpha = \sqrt{\frac{1}{2} \ln \frac{2nT}{\kappa}}$.

Algorithm 2 *SupLinUCB-S*

Initialization: $S \leftarrow \ln T, \Psi_{i,t}^{s} \leftarrow \phi$ for all $s \in [\ln T]$
for t = 1,2,..., T **do**
$s \leftarrow 1$ and $\hat{A}_1 \leftarrow \mathcal{N}$
$j \leftarrow 1 + (t \bmod n)$
repeat
Use *BaseLinUCB-S* with $\{\Psi_{i,t}^{s}\}_{i\in\mathcal{N}}$ and context vector x_t to calculate the width $w_{i,t}^{s}$ and upper confidence bound $ucb_{i,t}^{s} = (\hat{r}_{i,t}^{s} + w_{i,t}^{s})$, $\forall i \in \hat{A}_s$
if $j \in \hat{A}_s$ and $w_{j,t}^{s} > 2^{-s}$ **then**
Select $I_t = j$
Update the index sets at all levels:
$$\Psi_{i,t+1}^{s'} \leftarrow \begin{cases} \Psi_{i,t}^{s'} \cup \{t\} & \text{if } s = s' \\ \Psi_{i,t}^{s'} & \text{otherwise} \end{cases}$$
else if $w_{i,t}^{s} \leq \frac{1}{\sqrt{T}}, \forall i \in \hat{A}_s$ **then**
Select $I_t = argmax_{i\in\hat{A}_s} b_i \cdot (\hat{r}_{i,t}^{s} + w_{i,t}^{s})$
Update index sets at all levels for I_t:
$\Psi_{I_t,t+1}^{s'} \leftarrow \Psi_{I_t,t}^{s'}, \forall s' \in [S]$
else if $w_{i,t}^{s} \leq 2^{-s}, \forall i \in \hat{A}_s$ **then**
$\hat{A}_{s+1} \leftarrow \{i \in \hat{A}_s | b_i \cdot (\hat{r}_{i,t}^{s} + w_{i,t}^{s}) \geq \max_{a\in\hat{A}_s} b_a \cdot (\hat{r}_{a,t}^{s} + w_{a,t}^{s}) - 2^{1-s}\}$
$s \leftarrow s + 1$
else
Select $I_t = argmax_{i\in\hat{A}_s} b_i \cdot (\hat{r}_{i,t}^{s} + w_{i,t}^{s})$
until I_t is selected

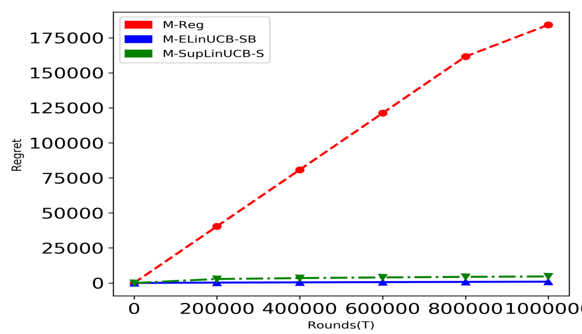

Figure 1: Regret vs Rounds (T)

Game Theoretic Analysis From the result in [3], an ex-post monotone allocation can be transformed to obtain a mechanism $\mathcal{M}$ such that $\mathcal{M}$ is EPIC and EPIR. As our proposed allocation rules *ELinUCB-SB* and *SupLinUCB-S* are ex-post monotone, we obtain EPIC and EPIR mechanism. All the details can be found in [1].

3 CONCLUSION

We believe that ours is the first attempt to design a non-exploration separated ConMAB mechanism. Although our mechanisms are randomized, they are game theoretically sound and scalable as compared to *M-Reg*. Further, in terms of regret, *M-ELinUCB-SB* and *M-SupLinUCB-S* outperforms *M-Reg* in experiments and theoretically *M-SupLinUCB-S* matches the regret in non-strategic setting up to a factor of $O(n)$ which is the price of truthfulness.

REFERENCES

[1] Kumar Abhishek, Shweta Jain, and Sujit Gujar. 2020. Designing Truthful Contextual Multi-Armed Bandits based Sponsored Search Auctions. (2020). arXiv:cs.GT/2002.11349

[2] Gagan Aggarwal, Ashish Goel, and Rajeev Motwani. 2006. Truthful Auctions for Pricing Search Keywords. In *Proceedings of the 7th ACM Conference on Electronic Commerce (EC '06)*. ACM, New York, NY, USA, 1–7. https://doi.org/10.1145/1134707.1134708

[3] Moshe Babaioff, Robert D. Kleinberg, and Aleksandrs Slivkins. 2015. Truthful Mechanisms with Implicit Payment Computation. *J. ACM* 62, 2, Article 10 (May 2015), 37 pages. https://doi.org/10.1145/2724705

[4] Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. 2009. Characterizing Truthful Multi-armed Bandit Mechanisms: Extended Abstract. In *Proceedings of the 10th ACM Conference on Electronic Commerce (EC '09)*. ACM, New York, NY, USA, 79–88. https://doi.org/10.1145/1566374.1566386

[5] Nikhil R. Devanur and Sham M. Kakade. 2009. The Price of Truthfulness for Pay-per-click Auctions. In *Proceedings of the 10th ACM Conference on Electronic Commerce (EC '09)*. ACM, New York, NY, USA, 99–106. https://doi.org/10.1145/1566374.1566388

[6] Nicola Gatti, Alessandro Lazaric, and Francesco Trovò. 2012. A Truthful Learning Mechanism for Contextual Multi-slot Sponsored Search Auctions with Externalities. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12)*. ACM, New York, NY, USA, 605–622. https://doi.org/10.1145/2229012.2229057

[7] Ganesh Ghalme, Shweta Jain, Sujit Gujar, and Y Narahari. 2017. Thompson sampling based mechanisms for stochastic multi-armed bandit problems. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. 87–95.

[8] Shweta Jain, Sujit Gujar, Satyanath Bhat, Onno Zoeter, and Y Narahari. 2018. A quality assuring, cost optimal multi-armed bandit mechanism for expertsourcing. *Artificial Intelligence* 254 (2018), 44–63.

[9] T Lai. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6 (1985), 4–22.

[10] Padala Manisha and Sujit Gujar. 2019. Thompson Sampling Based Multi-Armed-Bandit Mechanism Using Neural Networks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2111–2113.

[11] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani. 2007. *Algorithmic Game Theory*. Cambridge University Press, New York, NY, USA.

[12] Akash Das Sharma, Sujit Gujar, and Y Narahari. 2012. Truthful multi-armed bandit mechanisms for multi-slot sponsored search auctions. *Current Science* (2012), 1064–1077.