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 monotonic 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

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 $N = \{1, 2, \ldots, n\}$, 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 $M$ proceeds in discrete rounds $t = 1, 2, \ldots, T$. At each round $t$:

1. $M$ 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 $x_t$, $M$ chooses an agent $I_t \in N$.
3. $M$ observes $r_{I_t}$ which is 1 if it gets clicked and 0 otherwise. No feedback on the other agents.
4. $M$ determines payment $\beta_{I_t}$, $\beta_{I_t} \geq 0$ 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. $M$ 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 $\theta_i$. Thus CTR for agent $i$ at given round $t$ is: $\mu_i(x_t) = \mathbb{P}[r_{I_t}|x_t] = \theta_i^T x_t$. The objective of $M$ is to minimize social welfare regret which is given as:

$$\mathbb{E}_T(M) = \sum_{t=1}^{T} \left[ \theta_i^T x_t \cdot b_{I_t} - \theta_i^T x_t \cdot b_{I_t}^* \right]$$

(1)

Here, $i_t^*(x_t) = \arg \max_k \{b_{I_t} \cdot (\theta_k^T x_t)\}$.

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.
AAMAS 2020, May 9–13, Auckland, New Zealand

Extended Abstract

Algorithm 1 ElinUCB-SB

1. Inputs: n, T, α ∈ R+, bid vector b, batch size bs
2. Initialization: Sact = N, x′ ← 0_{d×1}, T′ = T
3. for all i ∈ N do
4. A_i ← I_i (d-dimensional identity matrix)
5. c_i ← 0_{d×1} (d-dimensional zero vector)
6. μ_k^i ̃b; μ_k^i ̃b = 0
7. for t = 1, 2, 3, ..., T′ do
8. I_t′ ← 1 + (t′ − 1) mod n
9. if I_t′ ∈ Sact then
10. for t = (t′ − 1)bs, ..., (t′ − bs − 1) do
11. Observe context as x_t
12. I_t ← I_t′
13. x′ ← ((t − 1)x′ + x_t)/t (averaging over contexts)
14. Observe click as r_t ∈ {0, 1}
15. A_{I_t} ← A_{I_t} + x_tx_t⊺, c_{I_t} ← c_{I_t} + r_t x_t, ̂θ_{I_t} ← A_{I_t}⁻¹c_{I_t}
16. if μ_k^i ̃b < μ_k^i then
17. (Ψ_k^i, r_k^i) ← (μ_k^i, ̂θ_k^i x_t) ⊆ (μ_i ̃b, x_t x_t⊺)
18. if max(μ_k^i ̃b, μ_k^i) < min(μ_k^i ̃b, μ_k^i) then
19. (μ_k^i ̃b, r_k^i) ← (max(μ_k^i ̃b, μ_k^i), min(μ_k^i ̃b, μ_k^i))
20. else
21. (μ_k^i ̃b, r_k^i) ← (μ_k^i ̃b, μ_k^i)
22. else
23. for t = (t′ − 1)bs, ..., (t′ − bs − 1) do
24. Observe x_t
25. I_t ← argmax_i b_i · (̂θ_i x_t), ⇒ I_t ∈ Sact
26. Observe click as r_t ∈ {0, 1}
27. for all agent i ∈ Sact do
28. if μ_k^i < max_{i ∈ Sact} μ_k^i then
29. Remove i from Sact

Intuition behind ElinUCB-SB: The algorithm maintains a set of active agents Sact. Once an agent is evicted from Sact, it cannot 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 ∈ Sact. If it evaluates to true the algorithm does exploration (lines[9-21]) else exploitation (lines[22-26]). 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 the round, elimination (lines[27-29]) is done which removes the agents j ∈ Sact from Sact if UCB of agent j is less than LCB of any other agent in Sact. 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√dT ln T) with probability at least 1 − κ if it is run with α = \sqrt{\frac{1}{2} \ln \frac{4\varepsilon}{\varepsilon}}.

Algorithm 2 SupLinUCB-S

1. Initialization: S ← ln T, Ψ_i^s ⊆ φ for all i ∈ [ln T]
2. for t = 1, ..., T do
3. s ← 1 and A_1 ← N
4. j ← 1 + (t mod n)
5. repeat
6. Use BaseLinUCB-S with \{Ψ_i^s\}_i∈N and context vector x_t to calculate the width w_i^s and upper confidence bound ucb_i^s = \langle \hat{c}_i^s + w_i^s \rangle, \forall i ∈ A_s
7. if j ∈ A_s and w_j > 2^{-s} then
8. Select I_t = j
9. Update the index sets at all levels:
10. \Psi_{I_t}^s ← \Psi_{I_t}^s \cup \{t\} if s = s'
11. else if w_j > 2^{-s} then
12. Select I_t = argmax_i b_i · (\hat{c}_i^s + w_i^s)
13. Update index sets at all levels for I_t:
14. \Psi_{I_t}^s ← \Psi_{I_t}^s \cup \{t\} \forall i ∈ [S]
15. else if w_j < 2^{-s} then
16. Select I_t = argmax_i b_i · (\hat{c}_i^s + w_i^s)
17. 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 M such that M is EPIC and EPR. As our proposed allocation rules ElinUCB-SB and SupLinUCB-S are ex-post monotone, we obtain EPIC and EPR 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


