A Dominant Strategy Truthful, Deterministic Multi-Armed Bandit Mechanism with Logarithmic Regret

(Extended Abstract)

Divya Padmanabhan*
IISc, Bangalore

Satyanath Bhat
IISc, Bangalore

Prabuchandran K. J.
IBM IRL, Bangalore

Shirish Shevade
IISc, Bangalore

Y. Narahari
IISc, Bangalore

1. INTRODUCTION

Multi-armed bandit (MAB) algorithms [3] are widely used in sequential decision making where the decisions are modeled as arms. Mechanism design has been applied in the context where the arms are controlled by *strategic* agents, leading to stochastic MAB mechanisms. An immediate example is sponsored search auctions (SSA). In SSA, there are several advertisers who wish to display their ads along with the search results generated in response to a query from an internet user. There are two components that are of interest to the planner or the search engine, (1) *stochastic component*: click through rate (CTR) of the ads or the probability that a displayed ad receives a click (2) *strategic component*: valuation of the agent for every click that the agent's ad receives. The search engine wants to allocate a slot to an ad which has the maximum social welfare (product of click through rate and valuation). However neither the CTRs nor the valuations of the agents are known. This calls for a learning algorithm to learn the stochastic component (CTR) as well as a mechanism to elicit the strategic component (valuation).

For single slot SSA, it is known that any truthful, deterministic MAB mechanism suffers a regret of $\Omega(T^{2/3})$ [2] where T is the time horizon. We observe that the characterization provided by Babaioff et al. [2] targets the worst case scenario. In particular, in the lower bound proof of $\Omega(T^{2/3})$, they consider an example scenario where the separation, $\bar{\Delta}$, between the expected rewards of the arms is a function of T. We note that when a similar example ($\bar{\Delta} = T^{-1}$) is used with the popular UCB algorithm [1], the number of pulls to the sub-optimal arm is linear, even in the non-strategic case. Hence, learning algorithms targeting such worst case scenarios are restrictive for a practical implementation, even when the arms are non-strategic. Motivated by this, our contributions are as follows.

Contributions

(1) We observe that in most MAB scenarios, the separation between the agents' rewards is rarely a function of T, and when the rewards of the arms are arbitrarily close, the regret contributed by such sub-optimal arms is negligible. We exploit this fact to allow the center to specify the resolution, Δ, with which the agents must be distinguished. We introduce the notion of Δ-Regret to formalize this regret.
(2) Using SSA as a concrete example, we propose a dominant strategy incentive compatible (DSIC) and individually rational (IR) MAB mechanism with a deterministic allocation and payment rule, based on ideas from the UCB family of MAB algorithms. The proposed mechanism Δ-UCB achieves a Δ-regret of $O(\log T)$.

2. THE MODEL: SINGLE SLOT SSA

Let $[K]$ be the set of agents or arms with cardinality K. Each of the K arms, when pulled, give rewards from distributions with unknown parameters. In SSA, the rewards of the arms correspond to clicks. The clicks for the advertisements are assumed to be generated from Bernoulli distributions with unknown parameters $\mu_1, \mu_2, \ldots, \mu_K$ where μ_i is the CTR of ad i. Our notations are provided in Table 1.

A mechanism $\mathcal{M} = \langle \mathcal{A}, P \rangle$ is a tuple containing an allocation rule $\mathcal{A}$ and a payment rule P. At every time step t, the allocation rule acts on a bid profile b of the agents as well as click realization ρ and allocates the slot to one of the K agents, say i. Then $\mathcal{A}(b, \rho, t) = i$. The payment rule $P^t = (P_1^t, P_2^t, \ldots, P_K^t)$. The allocation as well as payments in round t only depends on the click histories till t. The reader may refer to [2] for more details on click realization.

Let i_* be the arm with the largest social welfare, that is, $i_* = \arg\max_{i \in [K]} \{W_i \triangleq \mu_i v_i\}$, $W_* = \max_{i \in [K]} W_i$. We denote by I_t the agent chosen at time t as a shorthand for $\mathcal{A}(b, \rho, t)$. For any given $\Delta > 0$, define the set $S_\Delta = \{i \in [K] : W_* - W_i < \Delta\}$. S_Δ is the set of all agents separated from the best arm i_* with a social welfare less than Δ. Being indistinguishable, these arms contribute "zero" to the regret. The center fixes Δ based on the amount in dollars he is willing to tradeoff for choosing sub-optimal arms, given he has only a fixed time horizon T to his disposal. To capture this more practical notion of regret, we introduce the metric Δ-regret.

$$\Delta\text{-regret} = \sum_{t=1}^{T} (W_* - W_{I_t}) \mathbb{1}\left[I_t \in [K] \setminus S_\Delta\right] \qquad (1)$$

The center suffers a loss only when an agent with a social welfare greater than Δ away from W_* is chosen. Δ-regret captures this loss. The goal of our mechanism is to select agents to minimize the Δ-regret.

*Contact: divs1202@gmail.com

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

Symbol	Description
K, $[K]$	No. of agents and agent set
μ_i	CTR of agent i
v_i	Valuation of agent i for each click
W_i	Social welfare when agent i is allocated
$\rho_i(t)$	Click realization of agent i at time t
v_{max}	Maximum valuation over all agents
b_i	Bid of agent i
b	Bid profile of all agents
b_{-i}	Bid profile of all agents except agent i
$N_{i,t}$	No. of times agent i has been selected till time t
$\mathcal{A}(b,\rho,t)$	Allocation at time t for bid profile b and click realization ρ
i_*	Agent with maximum social welfare, ideally must be allocated at every time step
W_*	Social welfare when agent i_* is allocated
Δ	Input parameter by center to indicate the level at which the agents must be distinguished
S_Δ	Set of agents whose social welfare is less than Δ away from i_*. These agents do not contribute to Δ-regret.
$\widehat{\mu}^+_{i,t}$	UCB index corresponding to μ_i at time t
$\widehat{\mu}^-_{i,t}$	LCB index corresponding to μ_i at time t
$\widehat{\mu}_{i,t}$	Empirical CTR of agent i estimated from samples up to time t
P^t_i	Payment charged to agent i if he is allocated a slot at time t and he gets a click

Table 1: Notations for the single slot SSA setting

Algorithm 1 Δ-UCB Mechanism

Input:
T: Time horizon, K: number of agents
Δ : parameter fixed by the center
v_{max} : Maximum valuation of the agents

Elicit bids $b = (b_1, b_2, \ldots, b_K)$ from all the agents
Initialize $\widehat{\mu}_{i,0} = 0, N_{i,0} = 0\ \forall i \in [K]$
$u = 8Kv^2_{max} \log T/\Delta^2$
for $t = 1, \ldots, u$ **do** ▷ Exploration rounds
  $I_t = ((t-1) \mod K) + 1$ ▷ Round-robin exploration
  $N_{I_t,t} = N_{I_t,t-1} + 1$
  $\mathcal{A}(b,\rho,t) = I_t$ ▷ Allocate slot to agent I_t and observe $\rho_{I_t}(t)$
  $\widehat{\mu}_{I_t,t} = (\widehat{\mu}_{I_t,t-1}N_{I_t,t-1} + \rho_{I_t}(t))/N_{I_t,t}$
  $\epsilon_{I_t,t} = \sqrt{2\log T/N_{I_t,t}}$
  $\widehat{\mu}^+_{I_t,t} = \widehat{\mu}_{I_t,t} + \epsilon_{I_t,t}$, $\widehat{\mu}^-_{I_t,t} = \widehat{\mu}_{I_t,t} - \epsilon_{I_t,t}$
  $\widehat{\mu}^+_{i,t} = \widehat{\mu}^+_{i,t-1}, \widehat{\mu}^-_{i,t} = \widehat{\mu}^-_{i,t-1}\ \forall i \in [K] \setminus \{I_t\}$
  $P^t_i(b,\rho) = 0\ \forall i \in [K]$ ▷ Free rounds
end for
$\hat{i}_* = \arg\max_{i\in[K]} \widehat{\mu}^+_{i,u} b_i$
$j = \arg\max_{i\in[K]\setminus\{\hat{i}_*\}} \widehat{\mu}^+_{i,u} b_i$
$P = \widehat{\mu}^+_{j,u} b_j / \widehat{\mu}^+_{\hat{i}_*,u}$
for $t = u+1, \ldots, T$ **do** ▷ Exploitation rounds
  $\mathcal{A}(b,\rho,t) = \hat{i}_*$
  $P^t_{\hat{i}_*}(b,\rho) = P \times \rho_{\hat{i}_*}(t)$ ▷ Agent pays only for a click
  $P^t_i(b,\rho) = 0\ \forall i \in [K] \setminus \{\hat{i}_*\}$
  $\widehat{\mu}^+_{i,t} = \widehat{\mu}^+_{i,u}$, $\widehat{\mu}^-_{i,t} = \widehat{\mu}^-_{i,u}\ \forall i \in [K]$ ▷ No more learning
end for

3. OUR MECHANISM: Δ-UCB

The idea in our mechanism Δ-UCB is to explore all the arms in a round-robin fashion for a fixed number of rounds, without any payments from the agents. The number of exploration rounds is fixed based on the desired Δ, specified by the planner. At the end of exploration, with high probability, we are guaranteed that the arms not in S_Δ are well separated from the best arm i_* with respect to their social welfare estimates.

Further on, for all the remaining rounds, the best arm as per the UCB estimate of social welfare is chosen. However in the exploitation rounds, the chosen agent pays an amount for each click he receives. The amount to be paid by the agent is fixed based on the well known Vickrey Clark Grove (VCG) scheme [4]. Note that no learning place in these rounds and the UCB, LCB indices don't change thereafter. We present our mechanism in Algorithm 1.

4. PROPERTIES OF Δ-UCB

We now state the properties satisfied by Δ-UCB regarding truthfulness and regret. (Proofs are omitted due to space)

At any time step, every agent obtains some utility by participating in the mechanism. Let Θ_i denote the space of bids of agent i. Let $\Theta_{-i} = \Theta_1 \times \ldots, \times\Theta_{i-1} \times \Theta_{i+1} \times \ldots \times \Theta_K$ denote the space of bids of all agents other than agent i. We denote by $u_i(b_i, b_{-i}, \rho, t; v_i)$ the utility to agent i at time t when his bid is b_i, his valuation is v_i, the bid profile of the remaining agents is b_{-i} and the click realization is ρ. All agents are assumed to be rational and are interested in maximizing their own utilities.

In our setting the utility to an agent i is computed as,

$$u_i(b_i, b_{-i}, \rho, t; v_i) = (v_i - P^t_i(b,\rho))\mathcal{A}_i(b_i, b_{-i}, \rho, t)\rho_i(t) \quad (2)$$

Definition 1. *Dominant Strategy Incentive Compatible (DSIC) [2]: A mechanism $M = \langle \mathcal{A}, P\rangle$ is said to be dominant strategy incentive compatible if $\forall i \in [K], \forall b_i \in \Theta_i$, $\forall b_{-i} \in \Theta_{-i}, \forall \rho, \forall t, u_i(v_i, b_{-i}, \rho, t; v_i) \geq u_i(b_i, b_{-i}, \rho, t; v_i)$.*

Definition 2. *Individually Rational (IR): A mechanism $M = \langle \mathcal{A}, P\rangle$ is said to be individually rational if $\forall i \in [K]$, $\forall b_{-i} \in \Theta_{-i}, \forall \rho, \forall t, u_i(v_i, b_{-i}, \rho, t; v_i) \geq 0$.*

Theorem 3. *Δ-UCB mechanism is dominant strategy incentive compatible (DSIC) and individually rational (IR).*

Lemma 4. *Social Welfare UCB index: For an agent i, we define the social welfare UCB indices for agent i as,*

$$\widehat{W}^+_{i,t} = \widehat{\mu}_{i,t}v_i + \epsilon_{i,t}v_i = \widehat{\mu}_{i,t}v_i + \sqrt{2v_i^2 \log T/N_{i,t}} \quad (3)$$

$$\widehat{W}^-_{i,t} = \widehat{\mu}_{i,t}v_i - \epsilon_{i,t}v_i = \widehat{\mu}_{i,t}v_i - \sqrt{2v_i^2 \log T/N_{i,t}} \quad (4)$$

Then, $\forall t\ P\left(\left\{\omega : W_i \notin [\widehat{W}^-_{i,t}(\omega), \widehat{W}^+_{i,t}(\omega)]\right\}\right) \leq T^{-4}$.

Lemma 5. *For an agent i and time step t, let $B_{i,t}$ be the event $B_{i,t} = \{\omega : W_i \notin [\widehat{W}^-_{i,t}, \widehat{W}^+_{i,t}]\}$. Define the event $G = \bigcap_t \bigcap_{i\in[K]} B^c_{i,t}$, where $B^c_{i,t}$ is the complement of $B_{i,t}$. Then $P(G) \geq 1 - 1/T^2$.*

Theorem 6. *Suppose at time step t, $N_{j,t} > 8v^2_{max} \log T/\Delta^2$ $\forall j \in [K]$. Then $\forall i \in [K] \setminus S_\Delta$, $\widehat{W}^+_{i_*,t} > \widehat{W}^+_{i,t}$ with high probability ($= 1 - 2/T^4$).*

Theorem 7. *If the Δ-UCB mechanism is executed for a total time horizon of T rounds, it achieves an expected Δ-regret of $O(\log T)$.*

REFERENCES

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[2] M. Babaioff, Y. Sharma, and A. Slivkins. Characterizing truthful multi-armed bandit mechanisms. *SIAM Journal on Computing*, 43(1):194–230, 2014.

[3] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[4] W. Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1):8–37, 1961.