Thompson Sampling Based Mechanisms for Stochastic Multi-Armed Bandit Problems

Ganesh Ghalme
Indian Institute of Science
Bangalore, India
ganesh.ghalme@csa.iisc.ernet.in

Shweta Jain
Indian Institute of Science
Bangalore, India
jainshweta@csa.iisc.ernet.in

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

Y. Narahari
Indian Institute of Science
Bangalore, India
hari@csa.iisc.ernet.in

ABSTRACT

This paper explores Thompson sampling in the context of mechanism design for stochastic multi-armed bandit (MAB) problems. The setting is that of an MAB problem where the reward distribution of each arm consists of a stochastic component as well as a strategic component. Many existing MAB mechanisms use upper confidence bound (UCB) based algorithms for learning the parameters of the reward distribution. The randomized nature of Thompson sampling introduces certain unique, non-trivial challenges for mechanism design, which we address in this paper through a rigorous regret analysis. We first propose a MAB mechanism with deterministic payment rule, namely, TSM-D. We show that in TSM-D, the variance of agent utilities asymptotically approaches zero. However, the game theoretic properties satisfied by TSM-D (incentive compatibility and individual rationality with high probability) are rather weak. As our main contribution, we then propose the mechanism TSM-R, with randomized payment rule, and prove that TSM-R satisfies appropriate, adequate notions of incentive compatibility and individual rationality. For TSM-R, we also establish a theoretical upper bound on the variance in utilities of the agents. We further show, using simulations, that the variance in social welfare incurred by TSM-D or TSM-R is much lower when compared to that of existing UCB based mechanisms. We believe this paper establishes Thompson sampling as an attractive approach to be used in MAB mechanism design.

Keywords

Thompson Sampling, Multi-armed Bandit, Mechanism Design

1. INTRODUCTION

Consider the problem of a *center* or a *planner* that desires to obtain high quality service from a pool of service providers so as to minimize the total cost. The center's effective reward for procuring a service from a service provider consists of two components: (i) quality and (ii) cost of the service. In typical practical situations, the qualities of the service providers may be *unknown* to the center as well as to the service providers themselves. Moreover, the costs are private information of the service providers. Typical examples include crowdsourcing, online procurement, etc. The center can learn the qualities of the service providers by repeatedly requesting services from them and observing their performances over a given period of time.

If the agents (that is, the service providers) are honest in reporting their costs, the center can use *Multi-Armed Bandit* (MAB) algorithms to achieve a fine balance between exploration (obtaining repeated services from an agent so as to learn agent quality) and exploitation (selecting a best agent so far). However, a naïve implementation of MAB algorithms could fail when the agents are strategic and may misreport their costs to maximize their utilities. When the reward from each arm consists of a stochastic component as well as a strategic component, a principled approach for *MAB mechanism* design is much needed [7, 4].

A MAB mechanism learns the stochastic component of the reward (qualities in our case) and at the same time, ensures honest reporting of private information or strategic component (costs in our case) from the agents. There are two types of MAB mechanisms in the literature: (1) *Deterministic*, in which the payments are deterministic; however, these are known to incur high *regret* in social welfare [4, 7], and (2) *Randomized*, in which the payments are randomized; these achieve lower regret but at the cost of higher variance in utilities of agents [3, 5]. We aim to design MAB mechanisms with the dual objectives of reducing the regret in social welfare and reducing the variance in utility of agents. With this as the backdrop, we set for ourselves the following agenda.

Thompson Sampling for Allocation in MAB Mechanisms. Existing MAB mechanisms use frequentist approaches like upper confidence bound (UCB) [2] for learning the stochastic rewards of the agents [3, 7, 5]. In the current work, we follow the Bayesian approach for learning and propose the Thompson sampling algorithm to learn the rewards of strategic agents. Recent works have explored and analyzed the *Thompson sampling algorithm* [12] and have shown that Thompson sampling achieves slight better theoretical guarantees in terms of regret when compared to frequentist approaches (analytically and empirically) [10, 1, 6]. This, along with other properties such as robustness to delayed feedback, motivates the use of Thompson sampling based allocation. Our approach, to the best of our knowledge, is the first one to explore Thompson sampling in mechanism design context. On the flip side, the ran-

Appears in: *Proceedings 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.*

domized nature of Thompson sampling based allocation introduces certain unique challenges in mechanism design, which we effectively address.

Lower Regret through an Appropriate Notion of Truthfulness. Ex-post dominant strategy incentive compatibility (ex-post DSIC), the notion of truthfulness used in design of deterministic MAB mechanisms [4, 8] handles any manipulation by agents equipped with full knowledge of all future events. Though it is tempting to embrace such a strong notion of truthfulness, the approach imposes a severe restriction on the set of feasible, deterministic allocation rules and leads to high regret MAB mechanisms. In practical settings, the agents are unaware of future events, and thus, such a strong notion of truthfulness is not warranted. We therefore propose an appropriate and adequate variant of DSIC to achieve stronger guarantees on the regret.

Lower Variance in Utilities of Agents. To mitigate the problem of high regret of deterministic ex-post DSIC mechanisms, [3] proposed a class of randomized MAB mechanisms with randomized allocation and randomized payment rule. However, the agents face significant uncertainties in allocations as well as in payments. Thus, these randomized mechanisms suffer from a high variance in utilities of agents. A higher variance in utilities may lead to agents being uninterested in the mechanism. For example, crowd workers may not like high variance in rewards for the exact same service they offer. It is shown that this variance is inevitable if one desires to achieve ex-post DSIC [13]. The MAB mechanisms proposed in this paper achieve much lower variance in utilities of agents by working with our proposed notion of truthfulness and under realistic assumptions.

1.1 Contributions

We propose two complementary MAB mechanisms with allocations determined by the Thompson sampling approach. The first and immediate implication of using Thompson sampling is that these mechanisms achieve the same social welfare regret as achieved by the Thompson sampling algorithm for the classical MAB problem (Theorem 1).

Our first proposal is *TSM-D*, which has a deterministic payment rule. We prove the following properties of TSM-D: (a) the variance in agent utilities asymptotically tends to 0 (Theorem 2); (b) the mechanism is *ex-post individually rational with high probability* (Theorem 3); and (c) the mechanism is *within period dominant strategy incentive compatible with high probability* (Theorem 4).

Game theoretic properties satisfied by TSM-D are rather weak and to achieve stronger properties, we propose, as a key contribution in this paper, the mechanism *TSM-R*. TSM-R has a randomized payment rule. We prove the following properties of TSM-R: (a) with overlapping reward distributions, the variance in agent utilities asymptotically tends to $\frac{M^2}{6}$ (Theorem 5), where M is the reward a center gets if agent provides a service satisfactorily; (b) in the case of non-overlapping reward distributions, the variance in agent utilities is $\leq \frac{\Delta^2}{6\mu_{min}^2}$, where Δ is the difference in rewards to the center from the best agent and the worst agent and μ_{min} is the minimum quality of any agent (Theorem 6); (c) the mechanism is *ex-post individually rational* (Theorem 7); and (d) the mechanism is *within period DSIC* (Theorem 8), which is a weaker notion than ex-post DSIC but is much stronger than what is achieved by TSM-D.

We further obtain the following key additional insights via simulations: (a) the variances in agent utilities achieved by both TSM-D and TSM-R are significantly lower when compared to any randomized UCB algorithm based mechanisms, and (b) the convergence of variance of agent utilities in the proposed mechanisms is achieved in fewer rounds compared to current UCB based mechanisms.

2. THE MODEL

We address the following problem. There is a *center* that needs to procure a certain service repeatedly (say for T rounds) from a pool of agents $K = \{1, 2, \ldots, k\}$. Each agent i is characterized by two quantities: (i) quality $\mu_i \in [\mu_{min}, 1]$ with $\mu_{min} > 0$, the probability with which the center is satisfied with the service provided by an agent i and (ii) cost $c_i \in [c_{min}, c_{max}]$ for providing the service for one round. The center derives a utility of M for satisfactory service and a utility of 0 otherwise. Thus the welfare from an agent i is $r_i = M - c_i$ with probability μ_i and $-c_i$ with probability $1 - \mu_i$. If the qualities and costs of the agents are given, the welfare can be maximized by selecting an agent that maximizes the expected reward $R_i = M\mu_i - c_i$. Note that in our setting we maximize social welfare generated and not the total revenue from the agents. Many socially desirable outcomes like a government seeking help from the volunteers through crowdsourcing for disaster relief or seeking services for public projects require social welfare maximization.

The expected reward from an agent i consists of two components: (i) $M\mu_i$, that is unknown to the center as well as to the agents and is stochastic and (ii) $-c_i$, which is private to the agent i and is strategic. When the costs are known, the social welfare can be maximized using classical multi-armed bandit problem [11] where each agent i corresponds to an arm with reward r_i which is observed only when the agent i is selected. Let the history of allocations and observations about the agents' qualities till round $t-1$ be denoted by h_t which is a common knowledge.

As the agents are strategic, appropriate incentives should be offered to elicit their costs truthfully. Let $b_{i,t}$ denote the cost or bid reported by an agent i at round t. Let $b_{-i,t}$ be the bid vector of all the agents other than i and b_t denote the bid vector of all the agents at round t. Let $I_t \in K$ be the service providing agent selected at round t. The center pays $p_{i,t}(b_t; h_t)$ to the agent i if the agent is selected in round t. The utility of an agent i in round t is given by: $u_{i,t}(b_{i,t}, b_{-i,t}; h_t; c_i) = \mathbb{1}\{I_t(b_t; h_t) = i\}(p_{i,t}(b_t; h_t) - c_i)$. The expected social welfare regret over T rounds is given by: $\mathcal{R}_T = T \max_i\{M\mu_i - c_i\} - \sum_{t=1}^{T} \mathbb{E}_{I_t}[M\mu_{I_t} - c_{I_t}]$.

An *Allocation Rule* $\mathcal{A}$ takes a bid vector b_t and history h_t as inputs and outputs an index I_t of the selected agent at round t. A *Payment Rule* $\mathcal{P}$ determines the payment of each agent in each round. A MAB mechanism $\mathcal{M}$ is defined as $\mathcal{M} := (\mathcal{A}, \mathcal{P})$. When we use Thompson sampling based allocation, there is an inherent randomization caused by the sampling of the rewards for each agent. Let ω_t denote any randomness that may occur in the mechanism in round t. The notations used in the proofs are summarized in Table. 1. We aim at designing a MAB mechanism, which, in addition to ensuring truthful reporting by the agents, also enables the center to learn the qualities of the agents over multiple rounds by observing the quality of services of the selected agents.

2.1 Some Definitions

Based on the extent of information that the agents may have about the game and the way the agents decide on their strategies, we define different types of agents: omniscient, oblivious, myopic, and risk averse. The results in this paper make realistic assumptions on the types of agents and we state these assumptions as well.

DEFINITION 1. *(Omniscient Agent) We say an agent is Omniscient if at $t = 0$, an agent is aware of $\{\omega_t\}_{t=1}^{T}$. I.e. an agent has full a priori information about all the future events.*

T	Total number of rounds
$K = \{1, 2, \dots k\}$	Set of agents
$t \in \{1, \dots, T\}$	Running index on round
$\mu_i \in [\mu_{min}, 1]$	Quality of an agent i
$c_i \in [c_{\min}, c_{\max}]$	Cost per service of an agent i
c_{-i}	Cost vector of all agents other than i
M	Utility that center derives for satisfactorily service
r_i	Welfare or reward from agent i
$R_i = M\mu_i - c_i$	Expected welfare or reward from agent i
$b_t = \{b_{i,t}, b_{-i,t}\}$	Cost bid vector of all the agents for round t
I_t	Agent selected for round t
h_t	History of allocations and observed services till round t
$p_{i,t}(b_t; h_t)$	Payment to agent i for round t with history h_t and cost bid vector b_t
$u_{i,t}(b_t; h_t; c_i)$	Utility to agent i for round t with cost bid vector b_t, true cost c_i, and history h_t
$\mathcal{R}_T$	Expected social welfare regret
ω_t	Randomness in the mechanism in round t
$N_{i,t}$	Total number of services allocated to agent i till t
$\alpha_{i,t}$	Number of satisfactorily services provided by agent i till round t
$\beta_{i,t} = N_{i,t} - \alpha_{i,t}$	Number of non-satisfactorily services provided by agent i till round t
$\hat{\mu}_{i,t} = \frac{\alpha_{i,t}}{N_{i,t}}$	Empirical estimate of quality (μ_i) of an agent i
$\theta_{i,t} \sim Beta(\alpha_{i,t} + 1, \beta_{i,t} + 1)$	Sample drawn from Beta distribution with parameters $\alpha_{i,t} + 1$ and $\beta_{i,t} + 1$
$j_t^* = \arg\max_{i \neq I_t}\{M\theta_{i,t} - b_{i,t}\}$	Second best agent at round t based on sampled values
$\Delta_i = M\mu_1 - c_1 - M\mu_i - c_i$	Difference in expected reward of agent i from optimal agent
$\Delta = \max_i \Delta_i$	Maximum difference in expected reward of any sub-optimal agent from optimal agent
$E_t : \{\theta_{1,t} \geq \theta_{i,t} + \frac{c_1 - c_i}{M}, \forall i \neq 1\}$	Event that agent 1 is allocated at round t if all the agents bid truthfully
$q_t = \mathbb{P}(E_t)$	Probability with which event E_t occurs
$e_{i,t}(\gamma) = \sqrt{\frac{4\gamma \ln(t)}{N_{i,t}}}$	Exploration term for agent i at round t with parameter $\gamma \geq 1$

Table 1: Notation Table

Note that ex-post DSIC [4, 7] protects manipulations even from omniscient agents (with higher regret). However, in practice, the agents are not omniscient as they are unaware of the future events. Thus, it is adequate to consider the following types of the agents.

DEFINITION 2. *(Oblivious Agent) We say an agent is* oblivious *if at the beginning of the round t' the agent is not aware of T and is not aware of $\{\omega_t\}_{t \geq t'}$.*

Note that an oblivious agent is not aware about any future events and does not even know T. In many practical scenarios like crowdsourcing, agents are typically oblivious as there is no way an agent can be aware of how many service requests a center has. Further, since an agent completes a task successfully only with probability μ_i, the agent cannot be sure whether or not the center will be satisfied by the service provided by the agent. We further assume risk averse agents.

DEFINITION 3. *(Risk Averse Agent) An agent is said to be risk averse if the agent prefers a deterministic assured reward over a probability distribution having an expected value that is equal to the assured reward.*

We now define the notion of a myopic agent.

DEFINITION 4. *(Myopic Agent) We say an agent is* myopic *if the agent always maximizes the expected reward with respect to the current round and does not take into account any future rounds.*

Note: The myopic agents cannot manipulate the algorithm's learning of the stochastic component from future tasks as they are unaware of the time horizon (T). One can consider a more intricate model where an agent has a prior over the values of T and prevent possible manipulation accordingly. We wish to point out that agents which are risk averse and oblivious are also myopic.

2.1.1 Assumptions in the Model

(1) The agents are oblivious, that is, they do not have prior knowledge of T and do not have distributional knowledge of events in future rounds. This assumption, coupled with the assumption that the agents are risk averse, implies that the agents are myopic. A myopic agent in this setting assumes the current round to be the last round and does not manipulate based on the future events. A more sophisticated model would consider *foresighted agents* by working with a distribution over future utility gains. We do not consider foresighted agents in the current work. Note that a risk averse agent with distribution over future events may still prefer to lose in the current round by overbidding, if the expected future reward is higher. (2) The costs c_i's are constant throughout. However, we allow agents to update their bids $b_{i,t}$ at every round to impart flexibility to modify their bids based on the updated beliefs over their qualities.

2.2 Game Theoretic Properties

We now formally define key, relevant game theoretic properties. Some of these properties are motivated by realistic considerations regarding the types of agents that we deal with. We strive to work with realistic, appropriate notions of game theoretic properties rather than attempt to achieve strong properties that may well belong to the realm of impossibilities.

Incentive Compatibility

DEFINITION 5. *(**Within Period Dominant Strategy Incentive Compatible (WP-DSIC)**) We say a mechanism $\mathcal{M} = (\mathcal{A}, \mathcal{P})$ is* WP-DSIC*, if for all the agents and for all the rounds, the utility of an agent from truthful bidding is at least as much as the utility from any non-truthful bidding irrespective of the bids of other agents, i.e., $\forall i, \forall c_i, \forall t, \forall \omega_t, \forall h_t$ and $\forall b_{-i,t}$,*

$$u_{i,t}(c_i, b_{-i,t}; h_t; c_i|\omega_t) \geq u_{i,t}(b_{i,t}, b_{-i,t}; h_t; c_i|\omega_t),\ \forall b_{i,t}.$$

Note that WP-DSIC is weaker than ex-post DSIC. In WP-DSIC, truthful reporting of costs is weakly dominant strategy in the current round independent of history and any possible randomization. However, an omniscient agent may be able to manipulate WP-DSIC by losing in the current round so as to increase the utility in future rounds. Following is a weaker notion of incentive compatibility:

DEFINITION 6. ***(Within Period Dominant Strategy Incentive Compatible with High Probability (WP-DSICP))*** *We say a mechanism $\mathcal{M} = (\mathcal{A}, \mathcal{P})$ is* WP-DSICP *if for all the agents and for all the rounds, the probability of the utility of an agent from non-truthful bidding being more than the utility of an agent from truthful bidding asymptotically goes to* 0 *irrespective of the bids of other agents, i.e.,* $\forall i, \forall c_i, \forall t, \forall \omega_t, \forall h_t$ *and* $\forall b_{-i,t}$,

$$\mathbb{P}\left(u_{i,t}(c_i, b_{-i,t}; h_t; c_i|\omega_t) \leq u_{i,t}(b_{i,t}, b_{-i,t}; h_t; c_i|\omega_t)\right) \leq p_t,\ \forall b_{i,t},$$

with $\lim_{t\to\infty} p_t = 0$.

Individual Rationality

DEFINITION 7. ***(Ex-Post Individually Rational (EPIR))*** *We say a mechanism $\mathcal{M} = (\mathcal{A}, \mathcal{P})$ is* EPIR *if every agent has a non-negative utility with truthful bidding irrespective of the bids of other agents i.e.,* $\forall i, \forall c_i, \forall t, \forall h_t, \forall \omega_t$,

$$u_{i,t}(c_i, b_{-i,t}; h_t; c_i|\omega_t) \geq 0,\ \forall b_{-i,t}.$$

A weaker notion on individual rationality is defined as:

DEFINITION 8. ***(Ex-Post Individually Rational with High Probability (EPIRP))*** *We say a mechanism $\mathcal{M} = (\mathcal{A}, \mathcal{P})$ is* EPIRP *when the probability with which an agent gets negative utility with truthful bidding asymptotically goes to* 0 *irrespective of the bids of other agents, i.e.,* $\forall i, \forall c_i, \forall t, \forall h_t, \forall \omega_t$,

$$\mathbb{P}\left(u_{i,t}(c_i, b_{-i,t}; h_t; c_i|\omega_t) \leq 0\right) \leq p_t,\ \forall b_{-i,t},$$

with $\lim_{t\to\infty} p_t = 0$.

3. PROPOSED MECHANISMS: TSM-D AND TSM-R

First, we explain how Thompson sampling based allocation rule can be designed for our settings even when costs are known.

3.1 Thompson Sampling Based Allocation Rule

Thompson sampling algorithm maintains a distribution over rewards for each agent based on the observed rewards. At each round, the algorithm samples a reward for each agent with the current distribution and chooses an agent with the highest sampled reward.

In our setting, the welfare generated from each agent i is $M - c_i$ if agent i provides satisfactory service and $-c_i$ otherwise. We model this by assigning an independent Bernoulli random variable $X_{i,t}$ to each agent i for each round t such that $X_{i,t} = 1$ with probability μ_i. Thus, the reward at round t is $MX_{i,t} - c_i$ and the expected reward is $M\mu_i - c_i$. Note that $X_{i,t}$ captures the stochastic component of the reward. To adapt Thompson sampling for our problem, we maintain Beta priors on the stochastic rewards of agent i with parameters $\alpha_{i,t}$ and $\beta_{i,t}$. Here, $\alpha_{i,t}$ denotes the number of times the agent i has provided satisfactory service till round t and $\beta_{i,t}$ denotes the number of times the agent has failed to do so. A sample $\theta_{i,t}$ from this Beta distribution is then obtained and the agent I_t with maximum value of $M\theta_{i,t} - b_{i,t}$ is selected at round t. Once a service is procured from agent I_t, based on this agent's quality of service $X_{I_t,t}$, Beta priors for the next round are updated by appropriately updating the parameters $\alpha_{i,t+1}$ and $\beta_{i,t+1}$. Let $N_{i,t} = \alpha_{i,t} + \beta_{i,t}$ denote the number of times the agent i is selected for service till round t.

Without loss of generality, we assume that agent 1 is optimal, i.e., $M\mu_1 - c_1 \geq M\mu_i - c_i$, $\forall i$. Let $\Delta_i = M\mu_1 - c_1 - M\mu_i + c_i$ denote the difference in expected reward of agent i from the optimal agent and $\Delta = \max_i\{\Delta_i\}$. Thus, the expected social welfare regret can also be written as $\mathcal{R}_T = \sum_{i=2}^{k} \Delta_i \mathbb{E}[N_{i,t}]$.

Let $j_t^* = \arg\max_{i \neq I_t}\{M\theta_{i,t} - b_{i,t}\}$ be the second best agent at round t based on sampled values. For ease of notation, let $\hat{\mu}_{i,t} = \frac{\alpha_{i,t}}{\alpha_{i,t}+\beta_{i,t}}$ and event $E_t : \{\theta_{1,t} \geq \theta_{i,t} + \frac{c_1 - c_i}{M},\ \forall i \neq 1\}$. Note that E_t denotes the event that agent 1 is allocated at round t if all the agents bid truthfully. Let q_t be the probability of event E_t, i.e., $q_t = \mathbb{P}(\theta_{1,t} \geq \theta_{i,t} + \frac{c_1 - c_i}{M},\ \forall i \neq 1)$. Further, let $e_{i,t}(\gamma) = \sqrt{\frac{4\gamma \ln(t)}{N_{i,t}}}$ denote the exploration term for agent i at round t with parameter $\gamma \geq 1$. When it is clear from the context, we drop the parameters from utility function and denote the utility and expected utility of agent i at round t by $u_{i,t}(\cdot)$ and $U_{i,t}(\cdot)$ respectively.

We have the following Theorem that bounds the expected regret of the Thompson sampling based allocation rule for our setting:

THEOREM 1. *For any sub-optimal agent i, our allocation rule satisfies:* $\mathbb{E}[N_{i,T}] \leq O(\ln T)$*. Thus, expected regret* $\mathcal{R}_T = O(\ln(T))$.

PROOF. For the case of Bernoulli rewards, the proof of the above Theorem is given in [10]. However, in our setting, the rewards of the arms are $M - c_i$ with probability μ_i and $-c_i$ with probability $(1 - \mu_i)$. For this case also, the proof technique is similar to [10] with few changes. □

We now propose two mechanisms with different payment rules that could be used in conjunction with the above allocation rule.

3.2 MAB Mechanism: TSM-D

In order to achieve low variance in utilities, we first propose a naïve mechanism, TSM-D, with an *estimate based* payment rule (Algorithm 1). In each round, the payment to the selected agent is based on his *estimated externality* that is computed using the expected values ($\hat{\mu}_{i,t}$'s) of the Beta distributions maintained by the Thompson sampling algorithm with added exploration terms. Given the history of past allocation this is a *deterministic* quantity. These exploration terms are needed to ensure game theoretic

Algorithm 1: MECHANISM TSM-D

```
Input: Number of rounds $T$, Number of agents $k$, bids
        $\{b_{i,t}\}_{i=1}^{k}$ in each round $t \in \{1, 2, \dots, T\}$, Parameter $\gamma$
Output: Allocations $\mathcal{A} = \{I_t\}_{t=1}^{T}$ and payments
        $\mathcal{P} = \{p_{I_t,t}\}_{t=1}^{T}$.
Initialize: $\alpha_{i,1} = 0, \beta_{i,1} = 0\ \forall i \in \{1, 2, \dots, k\}$
for $t \leftarrow 1$ to $T$ do
    Sample: $\theta_{i,t} \sim Beta(\alpha_{i,t} + 1, \beta_{i,t} + 1)\ \forall i \in K$
    Allocate:
      $I_t = \arg\max_i\{M\theta_{i,t} - b_{i,t}\}$(break ties arbitarily)
    Payment: $p_{I_t,t} =$
      $M\hat{\mu}_{I_t,t} - M\hat{\mu}_{j_t^*,t} + b_{j_t^*,t} + 2M(e_{I_t,t}(\gamma) + e_{j_t^*,t}(\gamma))$
    Observe: The Bernoulli reward of an agent $I_t$ for round $t$
      i.e. $X_{I_t,t} = \begin{cases} 1 & \text{w.p. } \mu_{I_t} \\ 0 & \text{w.p. } 1 - \mu_{I_t} \end{cases}$
    Update:
    $\alpha_{I_t,t+1} = \alpha_{I_t,t} + \mathbb{1}\{X_{I_t,t} = 1\}$
    $\beta_{I_t,t+1} = \beta_{I_t,t} + \mathbb{1}\{X_{I_t,t} = 0\}$
    $\alpha_{i,t+1} = \alpha_{i,t},\ \beta_{i,t+1} = \beta_{i,t}\ \forall i \neq I_t$
```

properties. Since the parameters of the Beta distributions at the beginning of round t are fixed (based on h_t), the payment rule is deterministic. However, the mechanism is still a randomized one due to the randomization arising from Thompson sampling based allocation. The following Theorem holds for the variance in utilities of the agents.

THEOREM 2. *The variance in the utility of any agent i satisfies: $\lim_{t \to \infty} var(u_{i,t}(\cdot)) = 0$, when all the agents are truthful.*

PROOF. Let $C_{i,t} = -c_1 + M\hat{\mu}_{1,t} - M\hat{\mu}_{i,t} + c_i + 2M(e_{1,t} + e_{i,t})$ and $C = M + c_{max} + 4M\sqrt{6\gamma \ln(T)}$ (i.e.,$C_{i,t} \leq C, \ \forall i$). In TSM-D, for optimal agent, $u_{1,t}(\cdot)$ is $C_{j_t^*,t}$ if event E_t occurs and is 0 otherwise.

$$\begin{aligned}
\text{var}(u_{1,t}(\cdot)) &= \mathbb{E}[u_{1,t}(\cdot)^2] - \mathbb{E}[u_{1,t}(\cdot)]^2 \\
&= C^2_{j_t^*,t}\mathbb{P}(E_t) - C^2_{j_t^*,t}(\mathbb{P}(E_t))^2 \\
&= C^2_{j_t^*,t}\mathbb{P}(E_t)\mathbb{P}(E_t^c) \leq C^2_{j_t^*,t}\mathbb{P}(E_t^c) \leq C^2\mathbb{P}(E_t^c) \\
\sum_{t=1}^{T} \text{var}(u_{1,t}(\cdot)) &\leq C^2 \sum_{t=1}^{T} \mathbb{P}(E_t^c) \qquad \text{(summing over all rounds)} \\
&\leq C^2 \sum_{t=1}^{T} \mathbb{P}(\text{agent 1 is not selected in round } t) \\
&\leq C^2 \sum_{i=2}^{k} \mathbb{E}[N_{i,T}] \leq kC^2 O(\ln(T)) \\
&\qquad \text{(From Theorem 1)}
\end{aligned}$$

Thus, $\frac{1}{T}\sum_{t=1}^{T} \text{var}(u_{1,t}(\cdot)) \leq \frac{1}{T}kO((\ln(T))^2)$ as $C = O(\sqrt{\ln(T)})$ $\implies \lim_{t \to \infty} \text{var}(u_{1,t}(\cdot)) \to 0$. Similarly, for other agents $i \neq 1$, $\text{var}(u_{i,t}(\cdot)) \leq C^2\mathbb{P}(\text{agent } i \text{ is selected in round } t)$

$$\implies \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \text{var}(u_{i,t}(\cdot)) \leq \lim_{T \to \infty} \frac{C^2}{T}\mathbb{E}[N_{i,T}] \to 0. \quad \square$$

Note that low variance in utilities make our mechanism more useful as existing UCB based mechanisms suffer from huge variance which result in significant uncertainties to the agents. We now prove game theoretic properties of TSM-D. Before proving these properties, we need following propositions which can be proved using Hoeffding's inequality [9].

PROPOSITION 1. *For any agent i, we have:*

1. $\mathbb{P}(\hat{\mu}_{i,t} \leq \mu_i - e_{i,t}(\gamma)) \leq t^{-8\gamma}$
2. $\mathbb{P}(\hat{\mu}_{i,t} \geq \mu_i + e_{i,t}(\gamma)) \leq t^{-8\gamma}$

PROOF. Since, $\hat{\mu}_{i,t}$ is empirical mean from $N_{i,t}$ Bernoulli random variables with mean μ_i, from Hoeffding's inequality:

$$\mathbb{P}(\hat{\mu}_{i,t} \leq \mu_i - e_{i,t}(\gamma)) \leq exp\{-2N_{i,t}(e_{i,t}(\gamma))^2\} = t^{-8\gamma},$$
$$\mathbb{P}(\hat{\mu}_{i,t} \geq \mu_i + e_{i,t}(\gamma)) \leq exp\{-2N_{i,t}(e_{i,t}(\gamma))^2\} = t^{-8\gamma}$$

□

Part of the proof of our next proposition (Proposition 2) is provided in [10] which uses Beta-Bernoulli trick which is given as follows:

Beta-Bernoulli Trick: Let $F^{Beta}_{a,b}$ denote the cdf of a Beta distribution with parameters a and b and let $F^B_{j,\mu}$ (resp. $f^B_{j,\mu}$) the cdf (resp. pdf) of a Binomial distribution with parameters j and μ. Then:

$$F^{Beta}_{a,b}(y) = 1 - F^B_{a+b-1,y}(a-1) \qquad (1)$$

PROPOSITION 2. *For any agent i, we have:*

1. $\mathbb{P}(\theta_{i,t} \leq \mu_i - e_{i,t}(\gamma)) \leq t^{-2\gamma}$
2. $\mathbb{P}(\theta_{i,t} \geq \mu_i + e_{i,t}(\gamma)) \leq t^{-2\gamma}$

PROOF.

$$\begin{aligned}
1)\ \mathbb{P}(\theta_{i,t} \leq \mu_i - e_{i,t}(\gamma)) &= F^{Beta}_{\alpha_{i,t}+1,\beta_{i,t}+1}(\mu_i - e_{i,t}(\gamma)) \\
&= 1 - F^B_{\alpha_{i,t}+\beta_{i,t}+1,\mu_i - e_{i,t}(\gamma)}(\alpha_{i,t}) \\
&\qquad \text{(From Equation 1)}
\end{aligned}$$

Let $X_{1,l} \sim \text{Bernoulli}(\mu_i - e_{i,t}(\gamma))$ and $X_{2,l} \sim \text{Bernoulli}(\mu_i)$ be Bernoulli random variables. Further, let $Z_l = X_{2,l} - X_{1,l}$ be a discreet random variable with values $\{-1, 0, 1\}$ and mean $e_{i,t}(\gamma)$. Then,

$$\begin{aligned}
\mathbb{P}(\theta_{i,t} \leq \mu_i - e_{i,t}(\gamma)) &= 1 - \mathbb{P}\left(\sum_{l=1}^{N_{i,t}+1} X_{1,l} \leq \sum_{l=1}^{N_{i,t}} X_{2,l}\right) \\
&= \mathbb{P}\left(\sum_{l=1}^{N_{i,t}} X_{2,l} < \sum_{l=1}^{N_{i,t}+1} X_{1,l}\right) \\
&\leq \mathbb{P}\left(\sum_{l=1}^{N_{i,t}} Z_l < 1\right) = \mathbb{P}\left(\sum_{l=1}^{N_{i,t}} Z_l \leq 0\right) \\
&\qquad (X_{1,l} \leq 1\ \forall l) \\
&= \mathbb{P}\left(\sum_{l=1}^{N_{i,t}} (Z_l - e_{i,t}(\gamma)) \leq -N_{i,t}e_{i,t}(\gamma)\right) \\
&\leq exp\left\{\frac{-8\gamma N_{i,t}\ln(t)}{4N_{i,t}}\right\} \leq t^{-2\gamma} \\
&\qquad \text{(From Hoeffding's inequality)}
\end{aligned}$$

2) Following the similar steps:

$$\begin{aligned}
\mathbb{P}(\theta_{i,t} \geq \mu_i + e_{i,t}(\gamma)) &= 1 - F^{Beta}_{\alpha_{i,t}+1,\beta_{i,t}+1}(\mu_i + e_{i,t}(\gamma)) \\
&= F^B_{\alpha_{i,t}+\beta_{i,t}+1,\mu_i + e_{i,t}(\gamma)}(\alpha_{i,t}) \\
&\qquad \text{(From Equation 1)}
\end{aligned}$$

Let $X_{1,l} \sim \text{Bernoulli}(\mu_i + e_{i,t}(\gamma))$ and $X_{2,l} \sim \text{Bernoulli}(\mu_i)$. Let $Z_l = X_{1,l} - X_{2,l}$. Then:

$$\begin{aligned}
\mathbb{P}(\theta_{i,t} \geq \mu_i + e_{i,t}(\gamma)) &= \mathbb{P}\left(\sum_{l=1}^{N_{i,t}+1} X_{1,l} \leq \sum_{l=1}^{N_{i,t}} X_{2,l}\right) \\
&\leq \mathbb{P}\left(\sum_{l=1}^{N_{i,t}} Z_l \leq 0\right) \qquad (X_{1,l} \geq 0\ \forall l) \\
&= \mathbb{P}\left(\sum_{l=1}^{N_{i,t}} (Z_l - e_{i,t}(\gamma)) \leq -N_{i,t}e_{i,t}(\gamma)\right) \\
&\leq exp\left\{\frac{-8\gamma N_{i,t}\ln(t)}{4N_{i,t}}\right\} \leq t^{-2\gamma} \\
&\qquad \text{(From Hoeffding's inequality)}
\end{aligned}$$

□

For ease of presentation, let us denote the events, $F^1_{i,t} : \{\hat{\mu}_{i,t} \leq \mu_i - e_{i,t}(\gamma)\}$, $F^2_{i,t} : \{\hat{\mu}_{i,t} \geq \mu_i + e_{i,t}(\gamma)\}$, $F^3_{i,t} : \{\theta_{i,t} \leq \mu_i - e_{i,t}(\gamma)\}$ and $F^4_{i,t} : \{\theta_{i,t} \geq \mu_i + e_{i,t}(\gamma)\}$. From the above propositions, we have for any agent i, $\mathbb{P}(F^1_{i,t}) \leq t^{-8\gamma}$, $\mathbb{P}(F^2_{i,t}) \leq t^{-8\gamma}$, $\mathbb{P}(F^3_{i,t}) \leq t^{-2\gamma}$ and $\mathbb{P}(F^4_{i,t}) \leq t^{-2\gamma}$.

THEOREM 3. *TSM-D is EPIR with high probability (EPIRP).*

PROOF. To prove EPIR property of TSM-D, we show that the probability that any agent receives negative utility in round t is p_t with $\lim_{t\to\infty} p_t = 0$.

Since the utility of agents other than I_t is 0 at round t, it suffices to prove that the probability with which the agent I_t obtains negative utility is small. For ease of presentation, let $\bar{e}_t = e_{I_t,t}(\gamma) + e_{j_t^*,t}(\gamma)$, $\mu^-_{i,t} = \mu_{i,t} - e_{i,t}(\gamma)$, $\mu^+_{i,t} = \mu_{i,t} + e_{i,t}(\gamma)$. Similarly, let $\theta^-_{i,t} = \theta_{i,t} - e_{i,t}(\gamma)$, $\theta^+_{i,t} = \theta_{i,t} + e_{i,t}(\gamma)$. We will further use the following inequality: for any events F and G, $\mathbb{P}(F) = \mathbb{P}(F|G^c)\mathbb{P}(G^c) + \mathbb{P}(F|G)\mathbb{P}(G) \leq \mathbb{P}(F|G^c) + \mathbb{P}(G)$. Now, consider the probability that the agent I_t gets negative utility with truthful bidding:

$$
\begin{aligned}
&\mathbb{P}(u_{I_t,t}(c_{I_t}, b_{-i,t}; h_t; c_{I_t}) < 0)\\
&= \mathbb{P}\left(M\hat{\mu}_{I_t,t} - c_{I_t} < M\hat{\mu}_{j_t^*,t} - b_{j_t^*,t} - 2M\bar{e}_t\right)\\
&\leq \mathbb{P}\left(M\hat{\mu}_{I_t,t} - c_{I_t} < M\hat{\mu}_{j_t^*,t} - b_{j_t^*,t} - 2M\bar{e}_t | F^{1c}_{I_t,t}\right) + \mathbb{P}(F^1_{I_t,t})\\
&\hfill \text{(with } G = F^1_{I_t,t}\text{)}\\
&\leq \frac{\mathbb{P}\left(M\mu^-_{I_t,t} - c_{I_t} < M\hat{\mu}_{j_t^*,t} - b_{j_t^*,t} - 2M\bar{e}_t\right)}{1 - t^{-8\gamma}} + \mathbb{P}(F^1_{I_t,t})\\
&\leq \frac{\mathbb{P}(M\mu^-_{I_t,t} - c_{I_t} < M\mu^+_{j_t^*,t} - b_{j_t^*,t} - 2M\bar{e}_t)}{(1 - t^{-8\gamma})(1 - t^{-8\gamma})} + \mathbb{P}(F^1_{I_t,t}) + \mathbb{P}(F^2_{j_t^*,t})\\
&\hfill \text{(with } G = F^2_{j_t^*,t}\text{)}\\
&\leq \frac{\mathbb{P}(M\mu_{I_t} - c_{I_t} < M\mu_{j_t^*} - b_{j_t^*,t} - M\bar{e}_t)}{(1 - t^{-8\gamma})(1 - t^{-8\gamma})} + \mathbb{P}(F^1_{I_t,t}) + \mathbb{P}(F^2_{j_t^*,t})\\
&\hfill \text{(expanding } \mu^-_{I_t,t} \text{ and } \mu^+_{j_t^*,t} \text{ and rearranging)}\\
&\leq \frac{\mathbb{P}(M\theta^-_{I_t,t} - c_{I_t} < M\mu_{j_t^*} - b_{j_t^*,t} - M\bar{e}_t)}{(1 - t^{-8\gamma})(1 - t^{-8\gamma})(1 - t^{-2\gamma})} + \mathbb{P}(F^1_{I_t,t})\\
&+ \mathbb{P}(F^2_{j_t^*,t}) + \mathbb{P}(F^4_{I_t,t}) \hfill \text{(with } G = F^4_{I_t,t}\text{)}\\
&\leq \frac{\mathbb{P}(M\theta^-_{I_t,t} - c_{I_t} < M\theta^+_{j_t^*,t} - b_{j_t^*,t} - M\bar{e}_t)}{(1 - t^{-8\gamma})(1 - t^{-8\gamma})(1 - t^{-2\gamma})(1 - t^{-2\gamma})} + \mathbb{P}(F^1_{I_t,t}) + \mathbb{P}(F^2_{j_t^*,t})\\
&+ \mathbb{P}(F^4_{I_t,t}) + \mathbb{P}(F^3_{j_t^*,t}) \hfill \text{(with } G = F^3_{j_t^*,t}\text{)}\\
&\leq \frac{\mathbb{P}(M\theta_{I_t,t} - c_{I_t} < M\theta_{j_t^*,t} - b_{j_t^*,t})}{(1 - t^{-8\gamma})(1 - t^{-8\gamma})(1 - t^{-2\gamma})(1 - t^{-2\gamma})} + 2t^{-8\gamma} + 2t^{-2\gamma}\\
&\hfill \text{(from above propositions and expanding } \theta^-_{I_t,t}, \theta^+_{j_t^*,t}\text{)}\\
&= 2t^{-8\gamma} + 2t^{-2\gamma}
\end{aligned}
$$

Since I_t is selected at round t with true cost, $\mathbb{P}(M\theta_{I_t,t} - c_{I_t,t} < M\theta_{j_t^*,t} - b_{j_t^*,t}) = 0$. Hence, the last inequality follows. Since $\gamma \geq 1$, we get the desired result. □

THEOREM 4. *TSM-D is WP-DSIC with high probability (WP-DSICP).*

PROOF. We need to prove for any agent $i, \forall t, \forall \omega_t, \forall h_t, \forall b_{-i,t}$:

$$\mathbb{P}\left(u_{i,t}(c_i, b_{-i,t}; h_t; c_i|\omega_t) \leq u_{i,t}(b_{i,t}, b_{-i,t}; h_t; c_i|\omega_t)\right) \leq p_t, \ \forall b_{i,t}$$

with, $\lim_{t\to\infty} p_t = 0$
Consider two cases for fixed values of $\theta_{i,t}, \theta_{-i,t}$:
Case 1: $M\theta_{i,t} - c_i \leq \max_{j\neq i}\{M\theta_{j,t} - b_{j,t}\}$ **:** In this case, the utility of player i with bid $b_{i,t} \geq c_i$ is zero as $M\theta_{i,t} - b_{i,t} \leq M\theta_{j,t} - b_{j,t}$ for some $j \neq i$, and hence there is nothing to prove. However, if $b_{i,t} < c_i$ such that $M\theta_{i,t} - b_{i,t} > M\theta_{j,t} - b_{j,t}$ $\forall j$, then an agent i wins round t and hence we have, $I_t = i$. We now calculate the probability of agent i getting positive utility with such nontruthful bidding i.e.

$$\mathbb{P}\left(M\hat{\mu}_{j_t^\star,t} - b_{j_t^\star,t} < M\hat{\mu}_{i,t} - c_i + 2Me_{i,t}(\gamma) + 2Me_{j_t^\star,t}(\gamma)\right)$$

Using Proposition 1 and Proposition 2 and with arguments similar in Theorem 3 we have:

$$
\begin{aligned}
&\mathbb{P}\Big(M\hat{\mu}_{j_t^\star,t} - b_{j_t^\star,t} < M\hat{\mu}_{i,t} - c_i + 2Me_{i,t}(\gamma) + 2Me_{j_t^\star,t}(\gamma)\Big)\\
&\leq 2t^{-8\gamma} + 2t^{-2\gamma} = p_t\\
\implies\ &\mathbb{P}(u_{i,t}(b_{i,t}, b_{-i,t}; h_t; c_i|\omega_t) \geq 0) \leq p_t\\
\implies\ &\mathbb{P}(u_{i,t}(b_{i,t}, b_{-i,t}; h_t; c_i|\omega_t) \geq u_{i,t}(c_i, b_{-i,t}; h_t; c_i|\omega_t)) \leq p_t\\
&\hfill \text{as } u_{i,t}(c_i, b_{-i,t}; h_t; c_i|\omega_t) = 0
\end{aligned}
$$

Case 2: $M\theta_{i,t} - c_i > \max_{j\neq i}(M\theta_{j,t} - b_{j,t})$**:** In this case, when $M\theta_{i,t} - b_{i,t} > M\theta_{j,t} - b_{j,t}$ $\forall j$, then the utilities with bids $b_{i,t}$ and c_i are same. However, if $M\theta_{i,t} - b_{i,t} \leq M\theta_{j,t} - b_{j,t}$ for some j, then with bid $b_{i,t}$, agent i will get utility 0 and with bid c_i, the utility is positive with probability $1 - p_t = 1 - (2t^{-8\gamma} + 2t^{-2\gamma})$ from Theorem 3 and hence the inequality follows. □

Relation of γ with game theoretic properties: With high value of γ, game theoretic properties WPDSIC and EPIR are satisfied with high probability, however, the payment to an agent also increases. Thus, an appropriate value of γ is needed to have a good trade-off.

Though TSM-D satisfies WP-DSIC and EPIR with high probability, there is a small yet non-zero probability for the agent selected at round t obtaining a negative utility, leading to a possibility of misreporting of the costs. Thus, TSM-D is game theoretically a much weaker mechanism. This can be attributed to the fact that the Thompson sampling algorithm has inherent randomization in allocation, and the payment scheme is not exactly aligned with this randomization. This motivates our next main mechanism, TSM-R with stronger game theoretic properties but at the cost of a slightly higher variance.

3.3 MAB Mechanism: TSM-R

We now present TSM-R mechanism with Thompson sampling based allocation rule that always satisfies within period dominant strategy incentive compatible (WP-DSIC) and ex-post individual rationality (EPIR). The allocation rule of TSM-R is the same as TSM-D, but they differ in the payment rule. The payment rule in TSM-R is computed based on *sampled externality* in each round which is computed using sampled values ($\theta_{i,t}$'s), thus making payment rule in TSM-R a randomized mechanism. The payment to the selected agent I_t at round t is given by:

$$p_{I_t,t} = M\theta_{I_t,t} - M\theta_{j_t^*,t} + b_{j_t^*,t},$$

We now bound the variance in utility of agents in TSM-R using the following lemmas:

LEMMA 1. *Variance in utility of optimal agent (agent* 1*) satisfies:*

$$\lim_{t\to\infty} var(u_{1,t}(\cdot)) \leq \lim_{t\to\infty} \frac{M^2}{2} \max\left\{\frac{1}{N_{1,t}+3}, \frac{1}{N_{j_t^*,t}+3}\right\}.$$

PROOF. Let Y be a Bernoulli random variable with parameter q_t. From the allocation and payment rule of TSM-R, we have

$$u_{1,t}(\cdot) = \begin{cases} M\theta_{1,t} - M\theta_{j_t^*,t} + c_{j_t^*} - c_1, & \text{if } Y = 1\\ 0, & \text{otherwise.}\end{cases}$$

From the conditional variance formula:

$$\text{var}(u_{1,t}(\cdot)) = \mathbb{E}[\text{var}(u_{1,t}(\cdot)|Y)] + \text{var}(\mathbb{E}[u_{1,t}(\cdot)|Y])$$

We bound both the terms in RHS separately:

$$
\begin{aligned}
\text{var}(u_{1,t}(\cdot)|Y = 1) &= \text{var}(M\theta_{1,t} - M\theta_{j_t^*,t} + c_{j_t^*} - c_1)\\
&= M^2\text{var}(\theta_{1,t}) + M^2\text{var}(\theta_{j_t^*,t})\\
&\hfill (c_1, c_{j_t^*} \text{ being constant and } \theta_{1,t} \text{ and } \theta_{j_t^*,t} \text{ independent}^1)
\end{aligned}
$$

$$\begin{aligned}\text{var}(u_{1,t}(\cdot)|Y=0) &= 0\\ \mathbb{E}[\text{var}(u_{1,t}(\cdot)|Y)] &= M^2 q_t(\text{var}(\theta_{1,t}) + \text{var}(\theta_{j_t^*,t}))\\ &\leq 2M^2 \max\{\text{var}(\theta_{1,t}), \text{var}(\theta_{j_t^*,t})\}\\ &= 2M^2 \max\left\{ \frac{(\alpha_{1,t}+1)(\beta_{1,t}+1)}{(N_{1,t}+2)^2(N_{1,t}+3)},\right.\\ &\left.\frac{(\alpha_{j_t^*,t}+1)(\beta_{j_t^*,t}+1)}{(N_{j_t^*,t}+2)^2(N_{j_t^*,t}+3)}\right\}\\ &\qquad\text{(From the variance of Beta distribution)}\\ &\leq 2M^2 \max\left\{\frac{1}{4(N_{1,t}+3)}, \frac{1}{4(N_{j_t^*,t}+3)}\right\}\end{aligned}$$

The second term can be bounded by,

$$\mathbb{E}[u_{1,t}(\cdot)|Y] = \begin{cases}\Delta_{j_t^*} & \text{if } Y=1\\ 0 & \text{Otherwise}\end{cases}$$

$$\begin{aligned}\text{var}(\mathbb{E}[u_{1,t}(\cdot)|Y]) &= \mathbb{E}_Y[(\mathbb{E}[u_{1,t}(\cdot)|Y])^2] - (\mathbb{E}_Y[\mathbb{E}[u_{1,t}(\cdot)|Y]])^2\\ &= \Delta_{j_t^*}^2 q_t - (q_t\Delta_{j_t^*})^2 \leq \Delta^2 q_t(1-q_t)\end{aligned}$$

$$\begin{aligned}&\text{Thus, } \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\text{var}(\mathbb{E}[u_{1,t}(\cdot)|Y])\\ &\leq \frac{\Delta^2}{T}\lim_{T\to\infty}\sum_{t=1}^{T}q_t(1-q_t) \leq \frac{\Delta^2}{T}\lim_{T\to\infty}\sum_{t=1}^{T}(1-q_t)\\ &\leq \frac{\Delta^2}{T}\lim_{T\to\infty}\sum_{i=2}^{k}\mathbb{E}[N_{i,T}] \leq \Delta^2\lim_{T\to\infty}\frac{O(\ln(T))}{T} = 0.\end{aligned}$$

Thus,

$$\lim_{t\to\infty}\text{var}(u_{1,t}(\cdot)) \leq \lim_{t\to\infty}\frac{M^2}{2}\max\left\{\frac{1}{N_{1,t}+3}, \frac{1}{N_{j_t^*,t}+3}\right\}. \quad \square$$

LEMMA 2. *For any other agent $i \neq 1$, variance in utility asymptotically goes to 0.*

PROOF. Using similar arguments as in Lemma 1, one can show that for other agents $i \neq 1$:

$$\begin{aligned}\text{var}(u_{i,t}(\cdot)) &\leq \mathbb{P}(I_t = i)\left(2M^2\max_j \text{var}\left(\theta_{j,t} + \frac{\Delta^2}{4}\right)\right)\\ \implies \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\text{var}(u_{i,t}(\cdot)) &\leq \lim_{T\to\infty}\frac{D}{T}\mathbb{E}[N_{i,T}] \to 0.\end{aligned}$$

$D = 2M^2(\frac{1}{4} + \frac{\Delta^2}{4}) \geq 2M^2(\max_j \text{var}(\theta_{j,t} + \frac{\Delta^2}{4}))$, $\forall t$. □

THEOREM 5. *When the reward distributions are overlapping, the variance i utility of any agent i satisfies: $\lim_{t\to\infty} var(u_{i,t}(\cdot)) \leq \frac{M^2}{6}$, when all the agents are truthful.*

PROOF. The proof is immediate from Lemma 1 and Lemma 2 □

THEOREM 6. *When the reward distributions are non overlapping, then the variance in utility of optimal agent satisfies: $var(u_{1,t}(\cdot)) \leq \frac{\Delta^2}{6\mu_{min}^2}\ \forall t$*

PROOF. When the reward distributions are non overlapping, then we have: $\Delta_{j_t^*} > M(1+\mu_1-\mu_{j_t^*})$. Since $N_{1,t}, N_{j_t^*,t} \geq 0$ and, $\mu_{min} \leq \mu_1, \mu_{j_t^*} \leq 1$,
$\max\left\{\frac{M^2}{2(N_{1,t}+3)}, \frac{M^2}{2(N_{j_t^*,t}+3)}\right\} \leq \frac{\Delta_{j_t^*}^2}{6(1+\mu_1-\mu_{j_t^*})^2} \leq \frac{\Delta^2}{6\mu_{min}^2}$. Thus from Lemma 1, $\text{var}(u_{1,t}(\cdot)) \leq \frac{\Delta^2}{6\mu_{min}^2}$. □

Note: The above theorems provide an upper bound on the variance in agent utilities. However, through simulations we observed that the variance in agent utilities of TSM-R also approaches 0, though it is slightly higher than the variance obtained from that of TSM-D. We further make a note that when reward distributions are non-overlapping, then the regret is zero as the sub-optimal arms are never pulled in this case. Hence, sub-optimal arms' distributions are not learned by the mechanism. As optimal arms' payment is randomized with respect to sub-optimal arm we observe a non-zero variance in the utility of the optimal agent.
We now prove game theoretic properties satisfied by TSM-R.

THEOREM 7. *Mechanism TSM-R is EPIR.*

PROOF. From the payment rule given by TSM-R:

$$u_{i,t}(c_i, b_{-i,t}; h_t; c_i|\omega_t) = \mathbb{1}(I_t = i)(M\theta_{i,t} - M\theta_{j_t^*,t} + b_{j_t^*,t} - c_i).$$

It is enough to prove, $\forall t, \forall h_t,\ u_{I_t,t}(c_{I_t}, b_{-I_t,t}; h_t; c_{I_t}|\omega_t) \geq 0,\ \forall\omega_t$. As $I_t = \arg\max_i \{M\theta_{i,t} - b_{i,t}\}$, $M\theta_{I_t,t} - c_{I_t} \geq M\theta_{j_t^*,t} - b_{j_t^*,t}$, thus proving the inequality. □

THEOREM 8. *Mechanism TSM-R is WP-DSIC.*

PROOF. We need to prove $\forall t,\ \forall\omega_t,\ \forall h_t,\ \forall b_{-i,t}$:

$$u_{i,t}(c_i, b_{-i,t}; c_i; h_t|\omega_t) \geq u_{i,t}(b_{i,t}, b_{-i,t}; c_i; h_t|\omega_t),\ \forall b_{i,t}.$$

Consider the two cases for fixed values of $\theta_{i,t}, \theta_{-i,t}$:
Case 1) $\exists l \neq i,\ M\theta_{i,t} - c_i \leq M\theta_{l,t} - b_{l,t}$: In this case, if $M\theta_{i,t} - b_{i,t} \leq M\theta_{j,t} - b_{j,t}$ for some $j \neq i$, then the utility with bid c_i or $b_{i,t}$ is zero and hence there is nothing to prove. However, if $M\theta_{i,t} - b_{i,t} \geq M\theta_{j,t} - b_{j,t}\ \forall j$, then the utility of an agent i with bid $b_{i,t}$ is given by: $M\theta_{i,t} - c_i - M\theta_{j_t^*,t} + b_{j_t^*,t} \leq M\theta_{i,t} - c_i - M\theta_{l,t} + b_{l,t} \leq 0$. Whereas, with bid c_i, agent i would have got utility 0 and hence the inequality follows.
Case 2) $\forall j,\ M\theta_{i,t} - c_i \geq M\theta_{j,t} - b_{j,t}$: In this case, when $M\theta_{i,t} - b_{i,t} \geq M\theta_{j,t} - b_{j,t}\ \forall j$, then the utilities with bids $b_{i,t}$ and c_i are same. However, if $M\theta_{i,t} - b_{i,t} \leq M\theta_{j_t^*,t} - b_{j_t^*,t}$, then with bid $b_{i,t}$, agent i will get utility 0 and with bid c_i, the utility is positive and hence the inequality follows. □

4. ADDITIONAL INSIGHTS THROUGH SIMULATIONS

This section provides additional insights on the performance of the proposed mechanisms. Note that it has already been established empirically that Thompson sampling based allocation rule achieve lower regret as compared to frequentist based approaches [6]. We now show via simulation that TSM-D converges faster and at any given round exhibits a variance that is lower than that of TSM-R . We use a two agent setup with $M = 50$ with fixed agent qualities $\{0.87, 0.9\}$. The agent qualities are chosen as 0.87 and 0.9 to be close enough to ensure an overlap in the two reward distributions ($R_i's$) for any reasonable choice of Δ. Note that the behavior obtained through simulations with respect to the variance in utilities follows the same pattern for any choice of qualities as verified by our simulations. In order to capture randomization introduced by

[1] The distribution from which $\theta_{i,t}$'s are sampled are derived from observed rewards till round $t-1$. Hence, at round t, given these distribution, $\theta_{1,t}$ and $\theta_{j_t^*,t}$ are independent.

the Thompson sampling based algorithm, we fix the outcomes of the other events for all rounds, i.e. whether or not an agent provides satisfactory service. For each agent i at round t, the corresponding event is generated with a Bernoulli random variable with parameter μ_i. The mechanisms are simulated for 1000 iterations by fixing events, cost, and quality vector.

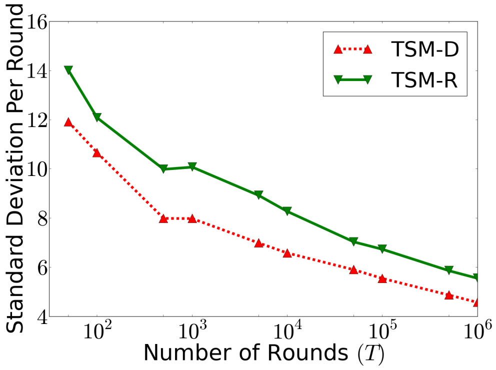

Figure 1: Standard deviation in utility of best agent with $\Delta = 5$

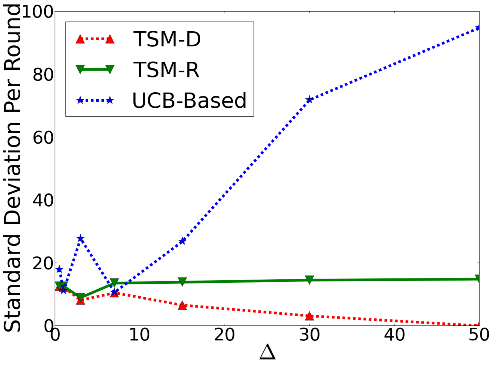

Figure 2: Standard deviation in utility of best agent for various values of Δ ($T = 5000$)

Figure 1 compares the variance in utilities of the best agent in TSM-D and TSM-R with the number of rounds for a fixed value of $\Delta = 5$. As expected, the variance in utility in TSM-R is slightly higher as compared to TSM-D. Note that TSM-R enjoys stronger game theoretic properties but at the cost of high variance. Figure 2 compares the variance in utilities of the best agent in TSM-D and TSM-R with the variance in utilities of the best agent in existing UCB-based mechanism [3] for various values of $\Delta \in [0.5, 50]$. We fixed $T = 5000$ for this experiment. The variance in utilities in TSM-D is the lowest between TSM-D, TSM-R, and UCB-Based. The variance in utilities of the best agent is significantly lower in our Thomson sampling based mechanism than the variance in utilities of the best agent in UCB-Based mechanism [3]. Figure 2 further shows for high values of Δ the variance of TSM-R converges to a constant value. This is due to the fact that for a higher value of Δ, the distributional overlap between the reward from the best agent and the second best vanishes. The high variance of UCB-based mechanism is attributed to a different resampling parameter that provides a trade-off between variance and the loss in social welfare. We used a reasonable value of this parameter (0.2) so as to avoid significant loss in social welfare.

	TSM-D	TSM-R	[3]
Allocation rule	Thompson sampling	Thompson sampling	UCB
Payment	Deterministic	Randomized	Randomized
EPIC	No	No	Yes
WPDSIC	with high probability	Yes	Yes
EPIR	with high probability	Yes	Yes
Variance in utility of optimal agent	0 (asymptotically)	0 (asymptotically) $\leq \frac{\Delta^2}{6\mu_{min}^2}$ (non-overlapping case)	much higher
Social welfare regret	logarithmic (lower constant)[6]	logarithmic (lower constant)[6]	logarithmic (higher constant)

Table 2: Properties satisfied by different mechanisms

Based on the analytical results in this paper and the simulation experiments, we now compare the proposed mechanisms alongside an existing, benchmark UCB based mechanism [3] on some of the important properties in Table 2. Table 2 clearly establishes that the Thompson sampling approach has certain characteristics that make it an attractive approach to be used in MAB mechanism design.

5. SUMMARY AND FUTURE WORK

This paper has explored Thompson sampling in the context of mechanism design for stochastic multi-armed bandit (MAB) problems. Many existing MAB mechanisms use upper confidence bound (UCB) based algorithms for learning the parameters of the reward distribution. Our motivation to use the Thompson sampling based approach has come from the known, superior regret performance of Thompson sampling when compared to UCB based approaches. The randomized nature of Thompson sampling introduces certain unique, non-trivial challenges for mechanism design, which we have effectively addressed in this paper. Our initial proposal was TSM-D, a MAB mechanism with deterministic payment rule. We showed that in TSM-D, the variance of agent utilities asymptotically approaches zero. However, the game theoretic properties satisfied by TSM-D (incentive compatibility and individual rationality with high probability) are rather weak. As one of our key contributions, we then proposed the mechanism TSM-R, with randomized payment rule, and proved that TSM-R satisfies appropriate, adequate notions of incentive compatibility and individual rationality. For TSM-R, we also established a theoretical upper bound on the variance in utilities of the agents. We further showed, using simulations, that the variance in social welfare incurred by TSM-D or TSM-R is much lower when compared to that of existing UCB based mechanisms. We believe this paper has established Thompson sampling as an attractive approach to be used in MAB mechanism design.

We analyzed only one deterministic payment rule. One can examine other deterministic payment schemes with stronger game theoretic properties. If deterministic payment schemes with stronger properties are not possible, a characterization result would be helpful. One can further extend this work to the case where the valuation of the agents dynamically changes over time. One can also relax the myopicity assumption about the worker and consider that each worker have a distributional knowledge about the future events.

REFERENCES

[1] S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. *CoRR*, abs/1111.1797, 2011.

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

[3] M. Babaioff, R. D. Kleinberg, and A. Slivkins. Truthful mechanisms with implicit payment computation. In *Eleventh ACM Conference on Electronic Commerce*, pages 43–52. ACM, 2010.

[4] M. Babaioff, Y. Sharma, and A. Slivkins. Characterizing truthful multi-armed bandit mechanisms: extended abstract. In *Tenth ACM Conference on Electronic Commerce*, pages 79–88. ACM, 2009.

[5] S. Bhat, S. Jain, S. Gujar, and Y. Narahari. An optimal bidimensional multi-armed bandit auction for multi-unit procurement. In *Fourteenth International Conference on Autonomous Agents and Multiagent Systems, AAMAS'15*, pages 1789–1790, 2015.

[6] O. Chapelle and L. Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2249–2257. Curran Associates, Inc., 2011.

[7] N. R. Devanur and S. M. Kakade. The price of truthfulness for pay-per-click auctions. In *Tenth ACM Conference on Electronic Commerce*, pages 99–106, 2009.

[8] N. Gatti, A. Lazaric, and F. Trovò. A truthful learning mechanism for contextual multi-slot sponsored search auctions with externalities. In *Thirteenth ACM Conference on Electronic Commerce*, pages 605–622, 2012.

[9] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[10] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory - 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings*, pages 199–213, 2012.

[11] T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4 – 22, 1985.

[12] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):pp. 285–294, 1933.

[13] C. A. Wilkens and B. Sivan. Single-call mechanisms. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 946–963, New York, NY, USA, 2012. ACM.