Mechanisms with Learning: Thompson Sampling based Mechanisms for Sleeping Multi-Armed Bandit Problems

Project Report Submitted in Partial Fulfillment of the Requirements

for the Degree of

Bachelor of Technology

in

Computer Science and Engineering

Submitted by

Akshay C. : 1510110035

Under the Supervision of

Dr. Y. Narahari

Department of Computer Science and Automation
Indian Institute of Science

SHIV NADAR UNIVERSITY

Department of Computer Science and Engineering

May 1, 2019

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Akshay C.

May 18, 2019

Computer Science and Automation
Indian Institute of Science

Y. NARAHARI, Ph.D.
FNA, FASc, FNAE, FNASc, FIEEE

April 27, 2019

This is to certify that C. Akshay, B.Tech, 4th year of Shiv Nadar University was an intern in the Game Theory Lab at the Department from January 1, 2019 to April 27, 2019.

The Project, titled "*Mechanisms with Learning: Thompson Sampling Based Mechanisms for Sleeping Multi-Armed Bandit Problems*" is a record of the bonafide work undertaken by him towards partial fulfillment of the requirements for the award of the degree of "Bachelor of Technology in Computer Science and Engineering".

I rate his performance as outstanding. I wish him all the best in his future pursuits.

Sincerely,

Dr. Y. NARAHARI,
PROFESSOR
Dept. of Computer Science and Automation
Indian Institute of Science
Bangalore - 562 012. (INDIA),
E-mail : hari@csa.iisc.ernet.in

Acknowledgement

I would like to thank Dr. Y. Narahari, IISc, for guiding me through this project. His mentorship and guidance has been very valuable.

I am grateful to all the lab members of the Game Theory Lab at IISc where this project was carried out.

I am also grateful to the faculty of the Department of Computer Science at Shiv Nadar University for these four wonderful years of learning.

Lastly I'd like to thank my parents who have continuously supported me.

Contents

List of Figures

List of Tables

Abstract

This thesis explores *Thompson Sampling* as a solution concept in the context of mechanism design for *Sleeping Multi-Armed Bandit* problems. The setting where the agents(arms) at play have both, a stochastic component as well as a strategic component to them. The stochastic component of the arms is learned through the use of Thompson Sampling while the strategic nature is handled through carefully constructed mechanism. While frequentist approaches such as UCB(Upper Confidence Bound) have been fairly well studied in the past Thompson Sampling has gained notoriety only recently. Some Previous known MAB mechanisms have made use Thompson Sampling to design mechanisms however the arms are assumed to be available throughout. This is not the case in many real world applications. This thesis extends previous known mechanism(TSM-R[2]) to a mechanism(TSSM-R) addressing the broader class of Sleeping Multi-Armed Bandit Problem. It also proves two key Game Theoretic properties which are *Ex-Post Individual Rationality* and *Within-Period Dominant Strategy Incentive Compatibility* for this new mechanism. This thesis also shows the superior performance of Thompson sampling approach to Sleeping MAB problems over frequentist approach like AUER(Awake Upper Estimated Reward) through simulations. It recreates empirically the previous known behaviour of Thompson Sampling where it adapts to the availability structure of arms better than AUER.

1 Introduction

The multi-armed bandit(MAB) problem is an extensively studied problem in online learning.

1.1 Motivation

Consider the case of crowdsourcing where there is a set of crowdworkers who are willing to do the work at different costs. Each has some probability of providing satisfactory work, this probability is known as his/her "quality". An entity wanting to repeatedly utilize their work will have to learn their qualities over time to obtain good work. This entity or "center" also has to incentivize the workers to reveal their true costs and not mis-report their costs i.e. they need to bid "truthfully".

There can be many other real world examples where the problem is the repeated selection rewards from multiple streams. Streams which may behave strategically and thus creating a need for the design of mechanisms to address this issue. Thus the whole setting needs to be studied in a Game Theoretic Framework.

Furthermore the assumption that all the arms will always be available is restricting. This needs to be relaxed and thus mechanisms need to designed in the context of Sleeping Multi-Armed Bandit problems.

1.2 Problem Characteristics

Having introduced the topic and provided the motivations for study, I now set forth exploring the problem with the following characteristics.

Learning the inherent stochasticity of arms: Evey agent has a stochastic component to its reward. The primary aim of all MAB algorithms is to learn this so as to minimize expected regret. In the case

of crowdsourcing this is the quality of the worker.

Strategic Nature of Arms: Every agent is strategic and will work towards maximizing his/her own utility. The mechanism is supposed to take care of this and incentivize the agents to reveal their true costs.

Thompson Sampling approach: Thompson Sampling has been shown to be very robust and performant approach of late and will be used to design mechanisms.

Sleeping Nature of Arms: One of the shortcomings of current MAB Mechanisms is that they assume that every arms is always available. This is certainly not true for most real world applications. The occasional non-availability of specific arms needs to be addressed and suitable mechanisms need to be designed. This is one of the goals that this thesis works towards.

2 Relevant Literature

2.1 Game Theory and Mechanism Design

Game Theory is the study of modelling the behaviour of *Rational* and *Intelligent* agents who strategically interact so as to maximize some utility in their own self-interest.

Rationality is the notion that each strategic agent has a certain well defined utility function and acts in a way to maximize it. Intelligence is the notion that the agent has the faculty to compute good strategy.

The foundations of Game Theory were laid down by John Von Neumann and Oskar Morgenstern in their 1944 book *Theory of Games and Economic Behavior*[11]. There are many other resources which have formed excellent references to the subject.

Mechanism Design is the "reverse engineering" of Game Theory. It is a branch of study which designs environments for strategic agents to interact in whereby the emergent selfish behaviour actually maximizes some Social Choice Function. Roger Myerson's book called Game Theory: Analysis of Conflict[12] serves as a reference to foundational results in Mechanism Design. Algorithmic Game Theory is a relatively new branch of study which aims to combine the disciplines of algorithm design and mechanism design. The book *Algorithmic Game Theory*[13] serves as the central reference.

2.2 Multi Armed Bandit Problem

The classical Multi Armed Bandit Problem was introduced by Robbins[14]. There are two approaches for designing algorithms for MAB prblem- Frequentist and Bayesian[1]. There have been a plethora of variants of the Multi-Armed-Bandit Problem which have been studied. For a review of

which one may look into Auer et al.[9]

2.2.1 Sleeping Varient

The sleeping Multi-Armed Bandit Problem was first studied by Kleinberg et al.[10] In this work the rewards are stochastic and there is no assumption on the structure of availability, it is adversarial.

2.2.2 Thompson Sampling

Thompson proposed a Bayesian algorithm in the 1933 paper[7] where he elaborated on the idea of assuming simple prior distribution on the parameters of the reward distributions. The lack of popularity of this approach can be attributed to the lack of theoretical guarantees of this approach.[1] There is a recent survey on this approach by Russo et al.[4]

2.3 Multi-Armed Bandit Mechanisms

This is a research direction which aims to study MAB problems in the presence of strategic agents. A review of all the advancements in this area can be found in the work of Jain et al.[3] This thesis is largly based on extending the work on MAB Mechanisms in the paper by Ganesh et al.[2]

3 The Model and the Mechanism

3.1 The Model

Here I describe the model of the problem. It is largely the same as the one in Ganesh et al.[2]. The key difference however being the fact that the availability of all agents in all round is no longer guaranteed. This is the precise idea that is captured by the Sleeping variant of the Stochastic Multi-Armed-Bandit problem.

I first describe the model as given in Ganesh et al.[2] and then I elaborate on how the model being considered here is different.

3.1.1 Model as given by Ganesh et al.[2]

The whole of this subsection appears in Ganes et al.[2] and is taken as is. The next section I highlight the extension to the model which characterizes the sleeping nature of agents. There is a *center* that needs to procure a certain service repeatedly (say for T rounds) from a pool of agents $K = \{1, 2, .., k\}$. Each agent i is characterized by two quantities: (i) quality $\mu_i \in [\mu_{min}, 1]$ with $\mu_m in > 0$, 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 $R_i = M\mu_i - c_i$.

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 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.

T	Time Horizon
K	Set of Agents
t	Current Round
μ_i	Quality of Agent i
c_i	Cost per service of agent i
c_{-i}	Cost vector of all agents apart from i
M	Utility the center derives from satisfactory service
r_i	Welfare or reward from agent i
$R_i = M\mu_i - c_i$	Expected welfare from agent i
b_t	Cost bid vector of all agents in round t
I_t	Agent selected for round t
h_t	History of allocations till round t
$p_{i,t}(b_t; h_t)$	Payment to agent i for round t
$u_{i,t}(b_t; h_t; c_i)$	Utility to agent i for round t with true cost c_i
R_t	Expected social welfare regret
$N_{i,t}$	Total number of services allocated to agent i till t
$\alpha_{i,t}$	Number of satisfactory services of i till t
$\beta_{i,t}$	Number of satisfactory services of i till t
$\theta_{i,t} \sim Beta(\alpha_{i,t}+1, \beta_{i,t}+1)$	Sample draws from Beta distribution
j_t^*	Second best agent at round t based on sampled values

Table 3.1: Table of Notations

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 $\mathcal{M} := (\mathcal{A}, \mathcal{P})$. The notations used in the proofs are summarized in the following Table. The aim is to design 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.

3.1.2 Some Definitions based on the information that the agents may have

DEFINITION 1. (Omniscient Agent) We say an agent is Omniscient if at $t = 0$ an agent has full apriori information about all the future events.[2]

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.[2]

DEFINITION 2. (Oblivious Agent) We say an agent is oblivious if at the beginning of the round $t = 0$ the agent is not aware of T and is not aware of any outcomes of randomizations in the mechanism. [2]

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.[2]

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.[2]

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.[2]

DEFINITION 5. (Awake) An agent i is said to be awake if he is available for allocation at round t.

DEFINITION 6. (Sleeping) An agent i is said to be awake if he is not available for allocation at round t.

3.1.3 Model assumptions and extensions to previous models

Ganesh et al. make the following model assumptions. (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.

Now I list the ways in which the current proposed model differs from the previous model. My model takes into consideration the possibility that the availability of all agents is not guaranteed at all rounds. Some of the agents may be *sleeping* and may not be available at certain rounds. Hence it is necessary to extend the previously known Mechanisms to take this situation into account. My model appropriately tweaks the previous mechanism and uses the Sleeping Multi-Armed Bandit problem model to model the situation. Thompson sampling is used to find near optimal allocations. I go on to prove two strong game theoretic properties with regard to this mechanism.

3.2 The Mechanism: TSSM-R

The TSM-R mechanism proposed by Ganesh et al.[2] which has strong Game Theoretic Properties has been extended to the class of problems involving the Sleeping variant of MAB. The new extended algorithm has been called TSSM-R and is given below. This differs from the previous algorithm in the fact that this now takes into consideration the availability of the agents before making an allocation.

3.2.1 Thompson Sampling approach to Allocation

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.[2]

However now prior to allocation we observe the availability of arms. We maintain Beta priors on the stochastic rewards of each arm(agent) and update the alpha and beta parameters of each arm appropriately as rounds progress. Here alpha denotes the number of times the agent has provided satisfactory service and beta denotes the number of times the agent has failed to do so. In the case of Bernoulli rewards the Expected regret satisfies $R_T = O(ln(T))$. The proof of this in the case of Bernoulli rewards is given in Chatterjee et al.[1]. In our setting the rewards are of the form $MX_{i,t} - c_i$. This also corresponds to a Bernoulli like situation where the rewards are linearly shifted. The bounds still holds for this approach as well. The reasoning is the same as that of Chatterjee et al.[1]

Given below is the modified version of TSM-D[2] called the TSSM-D

which works for the sleeping version of MAB.

Algorithm 1: Mechanism TSSM-R

Input: Number of rounds T, Number of agents k, bids $\{b_{i,t}\}_{i=1}^{k}$ in each round $t \in \{1, 2, \ldots, T\}$

Output: Allocations $\mathcal{A} = \{I_t\}_{t=1}^{T}$ and payments $\mathcal{P} = \{p_{I_t,t}\}_{t=1}^{T}$

Initialize: $\alpha_{i,1} = 0, \beta_{i,1} = 0 \forall i \in \{1, 2, \ldots, k\}$

for $t \leftarrow 1$ **to** T *do* **do**

Observe the subset of available arms $A_t \subseteq [K]$

Sample: $\theta_{i,t} \sim Beta(\alpha_{i,t} + 1, \beta_{i,t} + 1) \forall i \in A_t$

Allocate: $I_t = \text{argmax}_i\{M\theta_{i,t} - b_{i,t}\}$ (break ties arbitrarily)

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

Observe: The Bernoulli reward of an agent I_t for round t i.e.

$$X_{I_t,t} = \begin{cases} 1 & w.p. \mu_{I_t} \\ 0 & w.p. 1 - \mu_{I_t} \end{cases}$$

Update:

$\alpha_{I_t,t+1} = \alpha_{I_t,t} + \mathbf{1}\{X_{I_t,t} = 1\}$

$\beta_{I_t,t+1} = \beta_{I_t,t} + \mathbf{1}\{X_{I_t,t} = 0\}$

$\alpha_{i,t+1} = \alpha_{i,t}, \beta i, t+1 = \beta i, t, \forall i \neq I_t$

3.3 Game Theoretic Properties of TSSM-R

In this section I show that this randomized mechanism TSSM-R has two strong Game-Theoretic Properties. The first being Ex-post Individual Rationality and the second begin Within-Period Dominant Strategy Incentive Compatibility.

Individual Rationality is the property of a mechanism which ensures that no agent regrets having played the game. That is the utility of an agent is never negative for having participated in the mechanism. This is essential as it ensures that it is never a bad idea for an agent to play. In the worst case nothing may come out of it i.e. the utility may be zero but

he/she is never hurt from playing i.e the utility never drops to negative.

Within-Period Dominant Strategy Incentive Compatibility is the property of the mechanism which ensures that revealing their true costs i.e. bidding truthfully is a weakly dominant strategy of the agents. This essentially means that it is in the best interest of every agent to quote his true cost and not lie strategically with respect to his cost. This property ensures truthfulness among agents.

I prove that the mechanism TSSM-R possesses these two properties in the following two theorems. I do so by following similar arguments as in Ganesh et al.[2] The new mechanism TSSM-R is different from the older TSM-R mechanism in the the allocation part. The procedures post allocation are almost similar. Since these properties have to do more with what happens post allocation the proofs are structured very similarly to Ganesh et al.[2]

3.3.1 Theorem 1. TSSM-R is EPIR

According the model the utility an agent i receives in round t is given by

$$u_{i,t}(b_{i,t}; b_{-i,t}; h_t; c_t) = \mathbb{1}\{I_t(b_t; h_t) = i\}(p_{i,t}(b_t; h_t) - c_i).$$

The mechanism uses the payment rule($\mathcal{P}$) given by

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

Thus the utlity of an agent at round t is

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

To say that TSSM-R is EPIR we need to show that,

$$\forall t, \forall h_t, u_{I_t,t}(b_{i,t}; b_{-i,t}; h_t; c_t) \geq 0.$$

We know that $I_t = \text{argmax}_i\{M\theta_{i,t} - b_{i,t}\}$, and that j_t^* is the second best agent based on sampled values. Thus it is clear that

$$M\theta_{i,t} - c_{I_t} \geq M\theta_{j_t^*,t} - b_{j_t^*,t}.$$

This shows that the quantity $(M\theta_{i,t} - M\theta_{j_t^*,t} + b_{j_t^*,t} - c_i)$ is ≥ 0. This implies that the utility of an agent is always non-negative. This completes the proof. $\square$

3.3.2 Theorem 2. TSSM-R is WP-DSIC

To show that TSSM-R is Within-Period Dominant Strategy Incentive Compatible we need to prove that bidding truthfully is a very-wakly dominant strategy i.e.

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

We prove this case wise:

Case 1) There exists and agent $l \neq i$ such that $M\theta_{i,t} - c_i \leq M\theta_{l,t} - b_{l,t}$. This is a case where agent i is not the optimal agent in round t.

In this case we further divide into two cases

Case 1-a) When for some $j \neq i$ the following holds.

$$M\theta_{i,t} - b_{i,t} \leq M\theta_{j,t} - b_{j,t}$$

.

In this situation agent i isn't the one being allocated and thus his

utility is zero and hence there isn't anything to prove.

Case 1-b) However if

$$M\theta_{i,t} - b_{i,t} \geq M\theta_{j,t} - b_{j,t}$$

then the utility of agent i is given by

$$(M\theta_{i,t} - M\theta_{j_t^*,t} + b_{j_t^*,t} - c_i).$$

And this quantity is less than

$$(M\theta_{i,t} - M\theta_{l,t} + b_{l,t} - c_i) \leq 0.$$

This follows from our case assumption. Thus whereas with truthful bid of c_i the agent would have gotten zero utility he now has negative. Thus the inequality of WP-DSIC follows

Case 2) This is case where

$$\forall j, M\theta_{i,t} - c_i \geq M\theta_{j,t} - b_{j,t}$$

Case 2-a) In this case, when $M\theta_{i,t} - b_{i,t} \geq M\theta_{j,t} - b_{j,t}, \forall j$ the utilities with bids $b_{i,t}$ and c_i are same.

Case 2-b) However, if

$$M\theta_{i,t} - b_{i,t} \leq M\theta_{j_t^*,t} - b_{j_t^*,t}$$

then with a bid $b_{i,t}$ agent i will get utility of zero but a bid of c_i gets him positive utility and hence truthful bidding is again better. This proves that the inequality condition of WP-DSIC holds for all cases. Thus the proof is complete. □

4 Insights from Simulation

4.1 Simulation Parameters

This chapter is for empirically showing that an approach based on Thompson Sampling performs better than previous known frequentest algorithm AUER (Awake Upper Estimated Reward) when it comes to Regret Minimization. The methodology of the simulation closely follows that of Chatterjee et al[1]. The goal here is to show that the cumulative regret incurred by Thompson Sampling is much better than that of AUER. As described in Chatterjee et al.[1] we perform the simulation on synthetic datasets in which the parameters are chosen to highlight certain properties of the algorithms emperically.

In this simulation the following instance of Sleeping-MAB is considered.

- Number of Arms K is 10
- Time Horizon T is 10000
- Mean Rewards are as follows

$$\{0.5, 0.49, 0.48, 0.47, 0.46, 0.45, 0.44, 0.43, 0.42, 0.41\}$$

The motivation for the selection of these parameters is given by Chatterjee et al. in their paper[1] however Barto and Sutton in their book on reinforcement learning[8] provide a nice summary of why choosing 10 arms and time horizon of 10^4 is good enough. They call it *The 10 Arm Test Bed*

There are two very important settings in which the algorithm is tested and both are recreated in simulation in this chapter. These two settings

are as follows.

1. *The random availability setting*: Here each arm is independently available at any round with a probability of 0.5

2. *They changing suboptimal arm setting*: Here arm 1 is optimal and is always available. Along with with one other sub-optimal arm which is randomly chosen at each round is also awake. At any given round only these two arms are awake

4.2 AUER vs TS-SMAB– Performance comparison

In this section I show the results of the simulation

4.2.1 Random Availability Setting

In the random availability setting we do not observe a stark difference in performance however we can notice that the regret of Thompson sampling is lower

4.2.2 Changing Suboptimal Arm Setting

In this setting however we can see a very drastic increase in performance of the Thompson sampling algorithm as the algorithm performs better and is more adaptive to the structure of availability. We can see that the regret of Thompson sampling approach is much lower than AUER.

These simulation results agree very closely with the results of Chatterjee et al.[1] These results accurately recreate the empirical claims of Chatterjee et al.[1]. The adaptive nature of Thompson Sampling to the structure of availability of arms is a very strong observation which shows that the performance of Thompson Sampling approach to be strikingly better than AUER in certain conditions. This strange yet serendipitous

Figure 4.1: Comparing the regret of AUER and Thompson Sampling approach in the random availability setting, X axis- Time horizon, Y axis- Cumulative Regret

behaviour of this approach puts this approach as the best known solution concept for the Sleeping Multi-Armed Bandit Problem.

Figure 4.2: Comparing the regret of AUER and Thompson Sampling approach in the Changing Suboptimal Arm Setting, X axis- Time horizon, Y axis- Cumulative Regret

5 Summary and Future Work

5.1 Summary

This thesis started by introducing context in which the work is positioned, the rel event concepts were then elaborated. The relating and relevant literature was described. This thesis then described the previous model for MAB Mechanism and then introduced extensions to the model to take into consideration many real world scenarios where all the arms are not always available. It then proposed the mechanism(TSSM-R) with randomized allocation and payments based on Thompson Sampling which handles the Sleeping variant of MAB problem. It later proved two key game theoretic properties which are *Ex-Post Individual Rationality* and *Within-Period Dominant Strategy Incentive Compatibility* for this new mechanism. This thesis also shows the superior performance of Thompson sampling approach to Sleeping MAB problems over frequentist approach like AUER(Awake Upper Estimated Reward) through

simulations. It recreates empirically the previous known behaviour of Thompson Sampling where it adapts to the availability structure of arms better than AUER.

5.2 Directions for Future Work

.

1. Bounds on the variance in agent utilities: Since the mechanism is randomized there will be variation in the utilities of agents. There needs to be theoretical results on the bounds for this variance.

2. Multiple pull variants of MAB problems needs to looked further. Many real world scenarios rely on models where more than one arm needs to be pulled and there may be synergistic effects among certain combination of arms

3. Real world implementation of these algorithms needs to be done to see how these mechanisms fare out in the field.

4. Theoretical proof for a definite performance superiority of Thompson Sampling over Frequntist approaches needs to be produced. Currently there is only emperical evidence for this.

5. There is scope for studying Deterministic Payment scenarios which have been previously shown to have much lower variance on agent utilities.

6. There is scope for studies involving the relaxation on agent behaviour assumptions such as myopicity. The agents can have prior distributional knowledge of future events.

References

[1] Aritra Chatterjee, Ganesh Ghalme, Shweta Jain, Rohit Vaish, and Y Narahari. *Analysis of thompson sampling for stochastic sleeping bandits* In UAI, 2017.

[2] Ganesh Ghalme, Shweta Jain, Sujit Gujar, and Y. Narahari. *Thompson sampling based mechanisms for stochastic multi-armed bandit problems*. In Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS '17, pages 87–95, Richland, SC, 2017. International Foundation for Autonomous Agents and Multiagent Systems.

[3] Shweta Jain, Satyanath Bhat, Ganesh Ghalme, Divya Padmanabhan, and Y. Narahari. *Mechanisms with learning for stochastic multi-armed bandit problems* Indian Journal of Pure and Applied Mathematics, 47(2):229–272, Jun 2016.

[4] Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband and Zheng Wen, *A Tutorial on Thompson Sampling*, Foundations and Trends in Machine Learning: Vol. 11: No. 1, pp 1-96.

[5] Narahari, Yadati. *Game theory and mechanism design*. Vol. 4. World Scientific, 2014.

[6] Maschler, M., Solan, E., and Zamir, S. *Game Theory*. Cambridge University Press. 2013. doi:10.1017/CBO9780511794216

[7] Thompson W. R. *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples*. Biometrika. 1933 Dec 1;25(3/4):285-94.

[8] Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning (2nd ed.)*. MIT Press, Cambridge, MA, USA. 2018.

[9] Auer P, Cesa-Bianchi N, Fischer P. *Finite-time analysis of the multiarmed bandit problem*. Machine learning. 2002 May 1;47(2-3):235-56.

[10] Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. *Regret bounds for sleeping experts and bandits*. Mach. Learn. 80, 2-3 (September 2010), 245-272. 2010. DOI=http://dx.doi.org/10.1007/s10994-010-5178-7

[11] Von Neumann J, Morgenstern O. Theory of games and economic behavior (commemorative edition). Princeton university press; 2007 Mar 19.

[12] Myerson, Roger B. *Game Theory: Analysis of Conflict.* Harvard University Press, 1997.

[13] Nisan N, Roughgarden T, Tardos E, Vazirani VV, editors. *Algorithmic game theory.* Cambridge university press; 2007 Sep 24.

[14] Robbins H. *Some aspects of the sequential design of experiments.* Bulletin of the American Mathematical Society. 1952;58(5):527-35.