Thompson Sampling with Virtual Helping Agents

Kartik Anand Pant, Amod Hegde, and K. V. Srinivas, Member, IEEE

Abstract—We address the problem of online sequential decision making, i.e., balancing the trade-off between exploiting the current knowledge to maximize immediate performance and exploring the new information to gain long-term benefits using the multi-armed bandit framework. Thompson sampling is one of the heuristics for choosing actions that address this exploration-exploitation dilemma. We first propose a general framework that helps heuristically tune the exploration versus exploitation trade-off in Thompson sampling using multiple samples from the posterior distribution. Utilizing this framework, we propose two algorithms for the multi-armed bandit problem and provide theoretical bounds on the cumulative regret. Next, we demonstrate the empirical improvement in the cumulative regret performance of the proposed algorithm over Thompson Sampling. We also show the effectiveness of the proposed algorithm on real-world datasets. Contrary to the existing methods, our framework provides a mechanism to vary the amount of exploration/exploration based on the task at hand. Towards this end, we extend our framework for two additional problems, i.e., best arm identification and time-sensitive learning in bandits and compare our algorithm with existing methods.

Impact Statement—The Multi-Armed Bandit problem has been extensively studied in the last decade. It has wide-ranging applications from clinical trials to product assortment. Thompson sampling (TS) provides a simple heuristic solution to the MAB problem with sub-linear regret bounds. However, TS doesn’t provide control over the amount of exploration and exploitation executed by the policy. In this letter, we provide an extension to TS allowing variation in the exploration and exploitation in TS with strong theoretical guarantees. We believe that our framework can be easily augmented with the existing TS based solution and enhance its performance. Our work finds its application even in scenarios where the objective is to find satisfying (nearly optimal) actions on a short horizon or best actions over a long time horizon.

Index Terms—Multi-arm Bandits, Thompson sampling, Sequential Decision Making, Gaussian Bandits.

I. INTRODUCTION

N a stochastic multi-armed bandit (MAB) setting, an agent faces the problem of sequential decision making in the face of uncertainty. At each time step, the agent takes an action from a set of actions and each action produces a reward drawn from an underlying, fixed but unknown, distribution associated with that action. As the agent observes the reward at each time step, she learns about the underlying reward distributions and tries to optimize her long-term performance. The agent faces the dilemma of exploiting the already acquired knowledge to maximize her immediate rewards or exploring actions from which few/no observations have been made to acquire more knowledge for potential future gains while facing the risk of immediate loss.

Various algorithms have been proposed to solve the exploitation-exploration dilemma in the stochastic MAB problem. They include simple heuristics such as greedy and $\epsilon$-greedy algorithms [1], computationally intensive approaches such as Gittins indices [2], and the Upper Confidence Bound (UCB) family of algorithms which offer low computational cost and strong theoretical guarantees on the performance [3], [4], [5], [6], [7], [8].

Thompson proposed a simple heuristic for the stochastic MAB with Bernoulli rewards [9]. Starting with a prior distribution over the unknown parameters of the reward distribution of each action, the algorithm updates the posterior distributions as the actions are played. At each time step, an action is chosen according to its posterior probability of being the best action. This algorithm is known as Thompson sampling (TS) (and also as posterior sampling, probability matching) and has attracted a lot of attention in recent times. While [10], [11], [12], [13], [14] presented empirical studies showing excellent performance of TS in comparison with other state-of-the-art algorithms along with some weak theoretical guarantees of TS, [15], [16], [17], [18], [19] and [20] have presented rigorous theoretical analysis establishing tight bounds on the regret performance of TS.

In this paper, we present a modified TS algorithm, referred to as Thompson Sampling with Virtual Helping Agents and Combining (TS-VHA-C). The real (or, primary) agent playing the MAB game is assisted by $N − 1 > 0$ virtual helping agents, with each agent generating an independent sample from the posterior distribution of each arm: All the $N$ samples ($N − 1$ samples generated by the $N − 1$ virtual helping agents and the one generated by the primary agent), corresponding to each arm, are processed using a combiner and which arm to play next is decided based on the values of the combined samples. Here, we propose two linear combiners $C_1$ and $C_2$. Compared to the (conventional) TS, $C_1$ increases the exploitation at the expense of exploration and $C_2$ increases the exploitation at the expense of exploitation.

Importantly, our work may be considered as a framework for varying exploration vs. exploitation for Thompson sampling, by choosing the number of virtual helping agents and the type of combiner, enabling us to achieve a better regret performance (compared to TS) for some of the MAB problems. It is to be noted that one can design other combiners that achieve a...
different exploitation-exploration tradeoff.

Rest of the paper is organized as follows. After introducing the details of stochastic MAB problem and the Thompson sampling in Section II, we present the TS-VHA algorithm in Section III. Section IV states the main theoretical results that we present and the corresponding proofs. In Section V, we present simulation results to substantiate our theoretical results and Section VI concludes the paper.

II. THE STOCHASTIC MULTI-ARMED BANDIT PROBLEM

Consider an agent faced with a stochastic MAB problem. Given a slot machine with \( K \) arms, the agent has to choose an arm to play at each time step \( t \in \mathbb{Z}_{>0} \). The real-valued reward produced by each arm, when played, is a random variable whose distribution is fixed but unknown with a finite support over \([0, 1]\). The rewards obtained by playing an arm repeatedly are independent and identically distributed (i.i.d) and are independent of the plays of the other arms. The agent has to decide which arm to play at each time \( t \), based on its observations of the past \( t-1 \) plays and their outcomes, to maximize the expected total reward at time \( T \), a widely used performance metric in the stochastic MAB setting. The set of arms can also be referred to as the set of actions and playing arm \( i \) is equivalent to choosing action \( i \).

Denoting the (unknown) expected reward of arm \( i \) with \( \mu_i \) and the index of the arm played at time \( t \) with \( i(t) \), the expected total reward at time \( T \) is given by \( \mathbb{E} \left[ \sum_{t=1}^{T} \mu_{i(t)} \right] \). An equivalent and convenient metric to work with is the expected total regret, given by

\[
\mathbb{E}[R(T)] = \mathbb{E} \left[ \sum_{t=1}^{T} \mu^* - \mu_{i(t)} \right],
\]

where \( \mu^* := \max_i \mu_i \) and the expectation is over the random choices of arms played by the algorithm.

A. Thompson Sampling

As stated before, Thompson sampling takes a Bayesian approach. It starts by assuming an independent prior belief \( P(\mu_i) \) over the expected reward of each arm \( i \) and a likelihood function \( P(r \mid \mu_i) \) representing the probability of observing reward \( r \) upon playing arm \( i \). When an arm \( i \) is played, its posterior is updated based on the observed reward \( r \) using the Bayes rule: \( P(\mu_i \mid r) \propto P(r \mid \mu_i)P(\mu_i) \). At each time \( t \), an arm is played according to its posterior probability of having the highest mean reward; In practice, this is done by simply drawing a sample from the posterior distribution of each arm and playing the arm that produces the largest sample. Algorithm 1 presents the Thompson sampling.

III. THOMPSON SAMPLING WITH VIRTUAL HELPING AGENTS

Thompson sampling has three essential steps. First, the agent draws a sample from the posterior distribution of the expected reward of each arm, which acts as an estimate of the arm’s expected reward. Next, the agent selects the arm with the largest sample and observes a reward. Finally, the agent updates the posterior distribution of the expected reward of the selected arm based on the observed reward.

Our proposed algorithm introduces two significant changes to the Thompson sampling. First, we modify the sampling step of TS by employing \( N-1 \) virtual helping agents. Let \( \mathcal{K} = \{1, 2, \ldots, K\} \) represent the set of arms and let \( \mathcal{A} = \{1\} \cup \{2, \ldots, N\} = \{1, 2, \ldots, N\} \) denote the set of all agents, containing the real agent and the \( N-1 \) virtual helping agents. At every time step \( t \), all the \( N \) agents perform the sampling activity; i.e., every agent \( n \in \mathcal{A} \) draws a sample, independently, from \( P(\mu_i) \), the posterior distribution of the expected reward of \( i^{th} \) arm, for each \( i \) \in \mathcal{K} \). At the end of the sampling step, each agent \( n \in \mathcal{A} \) has \( K \) samples \( \theta_{i,n}(t), i = 1, \ldots, K \), where \( \theta_{i,n}(t) \) is the sample drawn by agent \( n \) from the posterior distribution of arm \( i \) at time \( t \). Note that the sampling activity (for generating the samples \( \theta_{i,n}(t), i = 1, \ldots, K, n = 1, \ldots, N \)) is independent across the agents and across the arms.

Next, for each arm \( i \in \mathcal{K} \), we combine the samples \( \theta_{i,n}(t), n = 1, \ldots, N \), using a combiner \( f : \mathbb{R}^{1 \times N} \rightarrow \mathbb{R} \) to arrive at the final combined estimate of the expected reward \( \theta_i(t) \) of the \( i^{th} \) arm. After the combining step, like in TS, we select the arm with the largest combined sample, observe the reward and update the posterior of the selected arm based on the observed reward. Note that the posterior update is same as that of the TS.

Algorithm 2 details the proposed Thompson sampling with virtual helping agents (TS-VHA). With no virtual agents (i.e., \( N-1 = 0 \)) and identity function as the combiner (i.e., \( f(\theta) = \theta \)), TS-VHA reduces to TS. In other words, TS-VHA can be interpreted as TS with \( N-1 \) virtual helping agents and a combiner. The \( N-1 \) virtual agents provide the real agent (who is actually trying to solve the MAB problem) with additional samples from the posterior to help her in manipulating the exploitation vs. exploration and deciding which arm to play at each time step. The agents are virtual as they do not really play the arms. Compared to TS, the additional cost of TS-VHA is in generating \( N-1 \) additional samples and processing them through the combiner function.

Algorithm 1  Thompson Sampling (TS)

Input: \( K \), priors \( P(\mu_i) \), likelihood \( P(r \mid \mu_i), i = 1, \ldots, K \).

for each \( t = 1, 2, \ldots \) do

Sample:

Draw \( \theta_i(t) \sim P(\mu_i), i = 1, \ldots, K \)

Select action:

Play arm \( i(t) = \arg \max_i \theta_i(t) \) and observe its reward \( r(t) \)

Update distribution:

\( P(\hat{\mu}_i(t)) \leftarrow P(\hat{\mu}_i(t) \mid r(t)), \)

where \( P(\hat{\mu}_i(t) \mid r(t)) \propto P(r(t) \mid \hat{\mu}_i(t))P(\hat{\mu}_i(t)) \)
end for
Algorithm 2 TS with Virtual Helping Agents (TS-VHA)

Input: $K$, $N$, priors $P(\hat{\mu}_i)$, likelihood $P(r \mid \mu_i)$, $i = 1, \ldots, K$, Combiner $f$.

for each $t = 1, 2, \ldots$ do

Sample:
for each $n = 1, 2, \ldots, N$ do

Draw $\theta_{i,n}(t) \sim P(\mu_i)$, $\forall i \in \mathcal{K}$
end for

Combine:
$\theta_i(t) = f(\theta_{i,1}(t), \ldots, \theta_{i,N}(t))$, $\forall i \in \mathcal{K}$

Select action:
Play arm $i(t) = \arg \max \theta_i(t)$ and observe its reward $r(t)$

Update distribution:
$P(\hat{\mu}_i(t)) \leftarrow P(\mu_i(t) \mid r(t))$,
where $P(\hat{\mu}_i(t) \mid r(t)) \propto P(r(t) \mid \hat{\mu}_i(t))P(\hat{\mu}_i(t))$
end for

respectively. Both the combiners are linear, having the generic form given below.

$$f(\theta_{i,1}(t), \ldots, \theta_{i,N}(t)) = \sum_{n=1}^{N} c_n \theta_{i,n}, \forall i \in \mathcal{K}.$$  \hspace{1cm} (2)

Combiners C1 and C2 differ only in the choice of the coefficients $c_n$ as described in the following.

A. Combiner C1: Increasing Exploitation

Combiner C1 is given by

$$c_n = \frac{1}{N}, \forall n \in \mathcal{A}.$$  \hspace{1cm} (3)

Employing combiner C1 in TS-VHA (which will be referred to as TS-VHA-C1 now onward) leads to higher exploitation as compared to TS. Observe that, for any arm $i$, the variance of the distribution of $\theta_i(t)$ is $\frac{1}{N}$ times that of the distribution of $\theta_{i,n}(t)$, $n = 1, \ldots, N$, whereas the mean remains the same. Thus, for each arm, the variance of the distribution of the combined sample is lower compared to the variance of the posterior distribution of that particular arm. Thus, TS-VHA-C1 places more confidence on its empirical estimates $\hat{\mu}_i(t)$, resulting in increased exploitation and lower exploration when compared with TS. As $N \to \infty$, $\theta_i(t) \to \hat{\mu}_i(t)$ and TS-VHA-C1 emulates greedy decision making.

B. Combiner C2: Increasing Exploration

Combiner C2 is for increasing the variance of the distribution of $\theta_i(t)$, the combined sample for arm $i$, compared to the posterior distribution of arm $i$ and is given by the following set of coefficients $c_n$:

When $N$ is an even integer,

$$c_n = \frac{1}{N} \left( \sqrt{\frac{N^2 + 1}{N}} \right)^{n+1}, \ n = 1, \ldots, N.$$  \hspace{1cm} (4)

When $N$ is an odd integer,

$$c_n = \begin{cases} \frac{1}{N} \left( \sqrt{\frac{N^2 + 1}{N}} \right)^{n+1}, & n = 1, \ldots, N-1, \\ \frac{1}{N}, & n = N. \end{cases}$$  \hspace{1cm} (5)

Observe that, for any arm $i$, the variance of the distribution of $\theta_i(t)$ is $N$ times that of the distribution of $\theta_{i,n}(t)$, $n = 1, \ldots, N$, whereas the mean remains the same. Due to this increase in the variance, TS-VHA with $C2$ as its combiner (which will be referred to as TS-VHA-C2 hereafter) places less confidence on its empirical estimates $\hat{\mu}_i(t)$ and leads to higher exploration as compared to TS. Increasing the number of agents in this case makes TS-VHA-C2 to over-explore.

Though our focus in this work is on the two linear combiners C1 and C2 (given by (2), (3), (4), (5)), note that TS-VHA provides a generic framework to manipulate exploitation vs. exploration. One can design other forms of combiners with desired exploitation-exploration tradeoff for a wide-range of MAB problems for which TS can be applied. To highlight this point, we present C3, a third combiner.

C. Combiner C3: Dynamic Exploitation

C3 is a non-linear combiner that computes the combined sample $\theta_i(t)$ for each arm $i \in \mathcal{K}$ as

$$\theta_i(t) = \max \left( \sum_{n=1}^{N(t)} \frac{1}{N(t)} \theta_{i,n}(t), \min_{j \in \mathcal{K}} \tilde{\mu}_j(t) \right), \forall i \in \mathcal{K}.$$  \hspace{1cm} (6)

where $\tilde{\mu}_j(t)$ is the observed empirical mean reward of arm $i$ at time $t$, $\theta_{i,n}(t)$ is the sample generated by agent $n$ for arm $i$ at time $t$ and $N(t)$ is the total number of agents (among which $N-1$ are the virtual helping agents) which is dynamically determined at each time step $t$ as,

$$N(t) = \left\lceil \max(1, t \cdot \tilde{\Delta}) \right\rceil.$$  \hspace{1cm} (7)

where $\tilde{\Delta} = \tilde{\mu}^{(1)}(t) - \tilde{\mu}^{(2)}(t)$, and $\tilde{\mu}^{(1)}(t)$, $\tilde{\mu}^{(2)}(t)$ are the largest and the second-largest values, respectively, in the set $\{\tilde{\mu}_1(t), \ldots, \tilde{\mu}_K(t)\}$.

In (6), the term $\sum_{n=1}^{N(t)} \frac{1}{N(t)} \theta_{i,n}(t)$ is similar to combiner C1 with $N(t)$ number of agents. Note that, here, the number of agents $N(t)$ is a function of time unlike in C1. As discussed previously, the value of $N(t)$ commands the exploitation-exploration trade-off. The expression for $N(t)$ in (7) is based on the following intuition:

- With time, we expect our best empirical arm to be the optimal arm with increasing confidence. Thus, tuning TS to increase exploitation with time could reduce the regret incurred at later time steps and improve its performance.
- If the difference in the empirical means of the top two candidate arms $\tilde{\Delta}$ is high, it may suggest that the best empirical arm is indeed the optimal arm. Increasing
exploitation in this case may help reduce the cumulative regret on average. On the other hand, if \( \hat{\Delta} \) is low, we may have to explore more to better discern the optimal arm. Hence, we set \( N(t) \propto \hat{\Delta} \).

Based on the above intuition\(^1\), we hypothesize that the number of agents to be deployed should be dependent on time \( t \) and \( \hat{\mu}^2(t) - \hat{\mu}^1(t) \). Next, inspired from [14] we apply a max operation (6). Note that, unlike OBS, we take the maximum of the averaged estimate and the minimum among the empirical means of all arms. We observe that this step yields superior empirical results.

In Section V, we provide simulation studies that prove the effectiveness of TS-VHA, for all the three combiners C1, C2 and C3, on Bernoulli bandits and Gaussian bandits. However, due to the analytical tractability of Gaussian distribution, we focus on Gaussian bandits while analyzing the regret performance of TS-VHA. We provide the mathematical regret analysis only for combiners C1 and C2 and leave the analysis of the non-linear combiner C3 for future work.

D. Gaussian Bandits

In the rest of the paper, we mainly focus on stochastic multi-armed bandits where the likelihood of the reward distributions are Gaussian; To be precise, the likelihood of the reward of \( r_i(t) \), reward from arm \( i \) at time \( t \), given parameter \( \mu_i(t) \), is given by \( \mathcal{N}(\mu_i(t), 1) \). Denote the arm played at time \( t \) with \( i(t) \) and the number of plays of arm \( i \) until (and including) \( t-1 \) with \( k_i(t) \). Define \( \hat{\mu}_i(t) := \frac{\sum_{r=1}^{t} r_i(r)}{k_i(t)} \) and \( \tilde{\mu}_i(1) := 0 \). With the Gaussian likelihood, it is convenient to use Gaussian priors. Consider \( \mathcal{N}\left(\hat{\mu}_i(t), \frac{1}{k_i(t)+1}\right) \) as the prior for \( \mu_i \) at time \( t \). When arm \( i \) is played at time \( t \), the posterior distribution for \( \mu_i \), by applying Bayes rule, turns out to be \( \mathcal{N}\left(\tilde{\mu}_i(t+1), \frac{1}{k_i(t+1)+1}\right) \).

By using the Gaussian priors and likelihoods in Algorithm 1, TS can be employed for Gaussian bandits and is referred to as TS using Gaussian priors.

E. TS-VHA using Gaussian Priors

For a Gaussian bandit, we can apply TS-VHA by using Gaussian priors and Gaussian likelihoods in Algorithm 2 resulting in Algorithm 3. The additional step of generating multiple samples and combining them in TS-VHA may alter the distribution of \( \theta_i(t) \) in Algorithm 2. However, with Gaussian distributions and a linear combiner \( f \) (such as the one given by (2)), the distribution of \( \theta_i(t) \) remains Gaussian.

In the next Section, we bound the finite time expected regret of TS-VHA-C1 and TS-VHA-C2 for Gaussian bandits. Note that we have investigated the performance of TS-VHA-C3 only through simulation experiments, presented in Section V. Understanding the theoretical implications of C3 would be interesting and we will consider it in our future work.

IV. Regret Analysis

For the finite time regret analysis presented in this section, we consider employing TS-VHA with \( N - 1 > 0 \) virtual helping agents for Gaussian bandits with reward distribution being finite support over \([0, 1]\). When we chose C1 as the combiner, the variance of the combined sample \( \theta_i(t), \forall i \in K \), gets scaled by \( 1/N \) compared to the variance of the posterior distribution of arm \( i, \forall i \in K \). With combiner C2, variance of \( \theta_i(t), \forall i \in K \), gets scaled by \( N \) compared to the variance of the posterior distribution of arm \( i, \forall i \in K \).

In both the cases, the mean of the combined sample for each arm is equal to the mean of the posterior distribution of the corresponding arm. Equivalently, at each time step \( t \), TS-VHA-C1 results in scaling the variance of \( \theta_i(t), \forall i \in K \) by a factor \( 1/N \) and TS-VHA-C2 scales the variance of \( \theta_i(t), \forall i \in K \) by \( N \), when compared to Thompson sampling.

To unify the regret analysis of TS-VHA-C1 and TS-VHA-C2, we introduce \( 1/\gamma \) as the factor that determines the variance scaling. Thus, \( \gamma > 1 \) corresponds to TS-VHA-C1, \( \gamma \in (0, 1) \) corresponds to TS-VHA-C2.

Theorem 1. For the K-armed stochastic bandit problem, Thompson sampling with virtual helping agents using Gaussian priors and with variance scaling factor \( \gamma \) has expected regret at time \( T \geq K \).

For \( \gamma \in (0, 4) \),

\[
\mathbb{E}[R(T)] \leq \sum_{i=2}^{K} \left( c_i^1 \ln T \Delta_i^2 + f_i^1(\beta, \gamma, \epsilon) \Delta_i + \frac{9.5}{\Delta_i} \right) \tag{8}
\]

For \( \gamma \geq 4 \),

\[
\mathbb{E}[R(T)] \leq \sum_{i=2}^{K} \left( c_i^1 \ln T \Delta_i^2 + \epsilon \left( \frac{T^{1+\epsilon-\frac{2\beta}{\gamma}} - 1}{\epsilon + \frac{2\beta}{\gamma}} + g(\epsilon) + 1 \right) \Delta_i + \frac{9.5}{\Delta_i} \right) \tag{9}
\]

Algorithm 3 TS-VHA using Gaussian Priors

| Input: \( K, N \), Set \( \mu_i(1) = 0 \forall i \in \text{priors} \) \( \mathcal{P}(\hat{\mu}_i) \), likelihood \( P(r \mid \hat{\mu}_i), i = 1, \ldots, K \), Combiner \( f \) |
| for each \( t = 1, 2, \ldots \) do |
| Sample: |
| for each \( n = 1, 2, \ldots, N \) do |
| Draw \( \theta_{i,n}(t) \sim P(\hat{\mu}_i), \forall i \in K \) |
| end for |
| Combine: |
| \( \theta_i(t) = f(\theta_{i,1}(t), \ldots, \theta_{i,N}(t)), \forall i \in K \) |
| Select action: |
| Play arm \( i(t) = \arg \max \theta_i(t) \) and observe its reward \( r(t) \) |
| Update distribution: |
| \( P(\hat{\mu}_{i(t)} \mid r(t)) \propto P(r(t) \mid \hat{\mu}_{i(t)}) P(\hat{\mu}_{i(t)}) \) |
| end for |

\(^1\)We would like to emphasize that these points are only our intuition and we do not have a mathematical justification

\(^2\)As in [16], we consider the single-parameter model where only the mean of the reward distribution is unknown. We do not consider the two-parameter model where both mean and variance of the reward distribution are unknown.
where $c_i^1 = \frac{2(H(\beta)+1)\Delta_i}{(y_i-\mu_i)^2}$, $f_i^1(\beta, \gamma, \epsilon) = c'(g(\epsilon) + \zeta(\frac{2\beta}{\gamma} - \epsilon)) + 1$ and $\beta, \epsilon, y_i, x_i, \Delta_i, c'$ are all constants at time $T \geq K$.

### A. Proof of Theorem 1

We adopt the notation and definitions from [20] and follow the same methodology as that of [20] in analyzing the finite cumulative regret achieved by TS-VHA using Gaussian priors when employed over a $K$ armed Gaussian bandit. Without loss of generality, we assume that $\mu^* = \mu_1 > \max_{i \neq 1} \mu_i$.

**Definition 1.** $i(t)$ denotes the arm played at time $t$, $k_i(t)$ denotes the number of plays of arm $i$ until, and including, time $t - 1$. $\hat{\mu}_i(t)$ denotes the empirical mean, given by $\hat{\mu}_i(t) = \frac{1}{k_i(t)+1} \sum_{\tau=1}^{t}\theta_i(\tau)$, where $\theta_i(\tau)$ denotes the reward observed from arm $i$ at time $t$ and $\hat{\mu}_i(t) = 0$ when $k_i(t) = 0$.

**Definition 2.** $\theta_{i,N}(t)$ denotes the $n$th sample generated, independently, from $\mathcal{N}(\hat{\mu}_i(t), \frac{1}{k_i(t)+1})$, the posterior distribution of arm $i$ at time $t$ and $\theta_i(t) = f(\theta_{i,1}, \ldots, \theta_{i,N})$.

**Definition 3.** For arm $i = 2, \ldots, K$, $x_i$ and $y_i$ denote thresholds such that $\mu_i < x_i < y_i < \mu_1$.

**Definition 4.** For $i = 2, \ldots, K$, $E_{i}^{\mu}(t)$ is the event $\hat{\mu}_i(t) \leq x_i$ and $E_{i}^{\theta}(t)$ is the event $\theta_i(t) \leq y_i$.

**Definition 5.** $\mathcal{F}_i = \{i(\tau), r_{i(\tau)}(\tau), \tau = 1, 2, \ldots, t\}$ is the history of arm play until time $t$, where $i(\tau)$ is the arm played at time $\tau$ and $r_{i(\tau)}(\tau)$ is the reward observed from arm $i(\tau)$ at time $\tau$. Define $\mathcal{F}_0 = \emptyset$. By definition, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \mathcal{F}_{T-1}$.

**Definition 6.** Define $p_{i,t}$ as the probability
\[
  p_{i,t} = \Pr(\theta_i(t) > y_i \mid \mathcal{F}_{t-1}).
\]

The expected total regret in time $T$ is given by
\[
  \mathbb{E}[R(T)] = \mathbb{E} \left[ \sum_{t=1}^{T} (\mu^* - \mu_{i(t)}) \right] = \sum_i \Delta_i \mathbb{E}[k_i(T)],
\]
where $\Delta_i = \mu^* - \mu_i$ and $\mu^* := \max_i \mu_i$. In order to bound the expected regret, we need to bound $\mathbb{E}[k_i(T)]$ for $i \neq 1$, which can be decomposed into three terms as follows:

\[
  \mathbb{E}[k_i(T)] = \sum_{t=1}^{T} \Pr(i(t) = i)
\]
\[
  = \sum_{t=1}^{T} \Pr(i(t) = i, E_{i}^{\mu}(t), E_{i}^{\theta}(t)) + \sum_{t=1}^{T} \Pr(i(t) = i, E_{i}^{\mu}(t), E_{i}^{\theta}(t)) + \sum_{t=1}^{T} \Pr(i(t) = i, E_{i}^{\mu}(t))
\]
\[
  \text{(A)}
\]
\[
  \text{(B)}
\]
\[
  \text{(C)}
\]

We will now consider the terms (A), (B) and (C) individually. In the following, we consider $x_i = \mu_i + \frac{\Delta_i}{3}$, $y_i = \mu_i - \frac{\Delta_i}{3}$ and let $L_i(T) = \frac{2T}{\gamma(y_i-x_i)^2}$.

### B. Term (A)

For $k \geq 1$, let $\tau_k$ be the time step at which the first arm is played for the $k$th time, and let $\tau_0 = 0$. Then, as shown in [20] (cf. Eqn. (4), Section 2.1 in [20]), for $i \neq 1$,
\[
  \sum_{t=1}^{T} \Pr(i(t) = i, E_{i}^{\mu}(t), E_{i}^{\theta}(t)) \leq \sum_{k=0}^{T-1} \frac{\left(1 - \frac{p_i(t, \tau_{k+1})}{p_i(t, \tau_k)}\right)}{p_i(t, \tau_{k+1})}.
\]
(12)

It is easy to verify that (12) does not get affected by the distribution of the sample $\theta_i(t), \forall i \in K$. Hence, (12) holds good for TS as well as TS-VHA. We utilize (12) to prove the following bound on term (A).

**Lemma 1.** For $\gamma \in (0, 4)$,
\[
  \sum_{t=1}^{T} \Pr(i(t) = i, E_{i}^{\mu}(t), E_{i}^{\theta}(t)) \leq H(\beta)L_i(T) + c' \left(g(\epsilon) + \zeta \left(\frac{2\beta}{\gamma} - \epsilon\right) + \frac{4}{\Delta_i^2}\right).
\]
(13)

For $\gamma \geq 4$,
\[
  \sum_{t=1}^{T} \Pr(i(t) = i, E_{i}^{\mu}(t), E_{i}^{\theta}(t)) \leq H(\beta)L_i(T) + c' \left(g(\epsilon) + \frac{T^{1+\epsilon} - \frac{8\epsilon}{\gamma} - 1}{1 + \epsilon - \frac{8\epsilon}{\gamma}} + \frac{4}{\Delta_i^2}\right).
\]
(14)

where $\beta \in [1, 2]$, $\epsilon > 0$ and $\zeta$ is the Riemann zeta function.

**Proof:** Please refer to Appendix B.

### C. Term (B)

**Lemma 2.** For $i \neq 1$,
\[
  \sum_{t=1}^{T} \Pr(i(t) = i, E_{i}^{\mu}(t), \overline{E_{i}^{\theta}(t)}) \leq L_i(T) + \frac{1}{\Delta_i^2}
\]

**Proof:** $\sum_{t=1}^{T} \Pr(i(t) = i, E_{i}^{\mu}(t), \overline{E_{i}^{\theta}(t)})$ can be subdivided into two parts based on the values of $k_i(T)$.

\[
  \sum_{t=1}^{T} \Pr(i(t) = i, E_{i}^{\mu}(t), \overline{E_{i}^{\theta}(t)}) = \sum_{t=1}^{T} \Pr(i(t) = i, E_{i}^{\mu}(t), k_i(T) \leq L_i(T), \overline{E_{i}^{\theta}(t)})
\]
\[
  + \sum_{t=1}^{T} \Pr(i(t) = i, E_{i}^{\mu}(t), k_i(T) > L_i(T), \overline{E_{i}^{\theta}(t)})
\]
(15)

The first term on the RHS of (15) is bounded by $\mathbb{E} \left[ \sum_{i=1}^{T} \left(i(t) = i, k_i(t) < L_i(T)\right) \right]$ which is upper bounded
This results in, 

\[ E_{\infty}^{\text{in}} \text{ holds valid as a proof for our Lemma 3.} \]

For illustrating the potential benefits of TS-VHA. In the following, Section III, identity function as the combiner.

Note that, \( \theta_i(t) \sim \mathcal{N} \left( \hat{\mu}_i(t), \frac{1}{\gamma(k_i(t)+1)} \right) \). Let \( \nu_i(t) \sim \mathcal{N} \left( x_i, \frac{1}{\gamma(k_i(t)+1)} \right) \). Then, as \( \hat{\mu}_i(t) \leq x_i \)

\[
\Pr(\theta_i(t) > y_i | k_i(t) > L_i(T), \hat{\mu}_i(t) \leq x_i, \mathcal{F}_{t-1}) \\
\leq \Pr(\nu_i(t) > y_i | k_i(t) > L_i(T), \hat{\mu}_i(t) \leq x_i, \mathcal{F}_{t-1}) \tag{17}
\]

Using Inequality 3, for any fixed \( k_i(t) > L_i(T) = \frac{2 \ln(T \Delta_i^2)}{\gamma(y_i - x_i)^2} \),

\[
\Pr(\nu_i(t) > y_i) \leq \frac{1}{2} e^{-\gamma(k_i(t)+1)(y_i-x_i)^2} \leq \frac{1}{2} e^{-\gamma L_i(T)(y_i-x_i)^2} \leq \frac{1}{T \Delta_i^2} \tag{18}
\]

This results in,

\[
\sum_{i=1}^{T} \Pr(\theta_i(t) > y_i | k_i(t) > L_i(T), \hat{\mu}_i(t) \leq x_i, \mathcal{F}_{t-1}) \leq \frac{1}{\Delta_i^2},
\]

bounding the second term on the RHS of (15) with \( \frac{1}{\Delta_i^2} \).

D. Term (C)

Term (C) denotes the probability of pulling the sub-optimal arm \( i \) when it is neither well estimated nor well sampled.

Lemma 3. For \( i \neq 1 \),

\[
\sum_{i=1}^{T} \Pr(\theta_i(t) > y_i) \leq \frac{1}{d_i(x_i, \mu_i)} \leq \frac{9}{2 \Delta_i^2} + 1.
\]

**Proof:** The proof for Lemma 3 follows from [20]. Since, the proof for the Lemma 3 doesn’t depend on the posterior distribution of the arms, the proof provided for Lemma 2.15 in [20] holds valid as a proof for our Lemma 3.

\[ \mathbb{E}[k_i(T)] \] can be bounded by substituting Lemma 1, 2 and 3 in (11) and using this bound on \( \mathbb{E}[k_i(T)] \) in (10) completes the proof of Theorem 1.

V. SIMULATION EXPERIMENTS

In this section, we present computational experiments that illustrate the potential benefits of TS-VHA. In the following, TS-VHA-C1-VAn and TS-VHA-C2-VAn denote TS-VHA with \( n \) virtual helping agents, with combiner C1 and C2, respectively. Note that, as per the notation introduced in Section III, \( N - 1 = n \) and TS corresponds to \( N = 1 \) with identity function as the combiner.

A. Gaussian Bandits

We evaluate the performance of TS-VHA-C1 and TS-VHA-C2 and compare it with TS for Gaussian bandits. First, we consider a 20 armed bandit problem with reward from arm \( i \) modeled as \( \mathcal{N}(\mu_i, 1) \), where the mean reward \( \mu_i \) is independently sampled from \( U[0, 1] \). Fig. 1(a) shows the cumulative regret over 10000 time steps, averaged over 1000 independently sampled problem instances. Fig. 1(b) corresponds to a second Gaussian bandit problem with 200 arms, keeping all the other details same as that of the 20 armed bandit.

For TS-VHA-C1-VAn (TS-VHA-C2-VAn), exploitation (exploration) increases with \( n \), as compared to TS. As can be observed from the plots, increasing exploitation through TS-VHA-C1 improves the regret performance. It should be noted that having more exploitation might turn out to be counter-productive. As discussed in Section III-A, as \( n \) grows to a higher value, TS-VHA-C1-VAn starts behaving like the greedy algorithm. Observe that, in Fig. 1(a), TS-VHA-C1-VAn accumulates more regret and performs poorly relative to TS-VHA-C1-VAn, \( n = 1, 2, 3 \). Fig. 2 shows the distribution of final cumulative regret at the end of 10000 time steps, over 1000 runs, for a Gaussian bandit with 20 arms. Reward from arm \( i \) is distributed as \( \mathcal{N}(\mu_i, 1) \), where \( \mu_i, i = 1, \ldots, 20 \), is chosen by sampling independently from \( U[0, 1] \) once at the beginning of the experiment and kept constant throughout the 1000 runs. TS-VHA-C1 has a higher variance in its final cumulative regret and is thus not suitable for risk-sensitive scenarios. On the other hand, TS-VHA-C2 results in a lower variance in its final cumulative regret and may be preferred in risk-averse applications.

B. Bernoulli Bandits

We now evaluate the performance of TS-VHA over Bernoulli bandits, i.e., bandit problems with Bernoulli distributed rewards and Beta distribution as the prior.

1) Bernoulli Bandit with Randomized Mean Rewards: Similar to the Gaussian bandits discussed above, we consider two Bernoulli bandit problems, one with 20 arms and the other with 200 arms, with mean reward of each arm is independently sampled from \( U[0, 1] \).

Fig. 3(a) and Fig. 3(b) shows the cumulative regret over 100000 time steps, averaged over 1000 independently sampled problem instances, for the 20 armed bandit and the 200 armed bandit, respectively.

2) Real World Datasets: Here, we show the effectiveness of TS-VHA-C1 on the real-world data sets Coupon-Purchase [25] and edx-Courses [26].

The Coupon-Purchase dataset contains discount coupons applied to online purchases. From the dataset, we have considered only 142 coupons that correspond to products priced less than or equal to 200 price units and purchased by at least one customer (as in [27]). For these 142 coupons, we have extracted the purchase rate that lies within \([0, 0.3]\) and the final selling price normalized by 200 price units, which lies within \([0, 1]\). With each coupon as an independent arm that (when played) generates a binary valued reward according to a Bernoulli distribution, we formulate two bandit problems.
In the first one, the mean reward of an arm is equal to the corresponding coupon purchase rate and, in the second problem, the mean reward of each arm is equal to the coupon purchase rate multiplied by the corresponding selling price normalized by 200. By modeling the mean reward of each arm using a Beta distribution, we present the performance of TS and TS-VHA-C1 corresponding to the first and the second problem, respectively.

The edX-Courses dataset contains information regarding 290 Harvard and MIT courses and, as in [28], we compute the normalized course participation rates by dividing the number of certified participants in each course and obtain the course certification rates by re-using near-optimal satisficing arms rather than exploring unused arms. With per period regret as the performance metric of each arm, Fig. 5(a) and Fig. 5(b) compare the cumulative regret performance of TS and and TS-VHA-C1-VA\textsubscript{n}, \( n = 1, 2, 3 \), for the first and the second problem, respectively.

C. Time-Sensitive Bandit Learning

Most of the bandit algorithms focus on learning the optimal arm (or, action). Often, especially for bandit problems having a very large set of arms, convergence to optimality may take a long time rendering them not useful in some practical applications. For example, in the case of a recommender system, the learning agent may be required to impress upon the users through its near optimal recommendations during the early interactions; Or, the learning agent may not have enough number of interactions with each user to converge onto perfect recommendations.

In [29], the authors have addressed the problem of learning near-optimal satisficing actions considering situations where the near term performance is more important than the performance over an asymptotically long time horizon, or, the optimal action is costly to learn relative to near-optimal actions. Satisficing Thompson Sampling (STS), proposed in [29], performs time-sensitive learning by modifying the Select action step of TS.

Recall that, in TS (i.e., Algorithm 1), \( \theta_i(\tau) \) is the sample drawn from posterior of arm \( i \) at time \( \tau \), \( i(\tau) \) is the index of arm played at time \( \tau \) and \( \theta_{i(\tau)} \) is the expected reward of the arm played at time \( \tau \). At each time step \( t \), STS identifies an \( \epsilon \)-optimal action through the following Select action step.

**Select action (in STS):**

Let \( i(t) = \arg\max_i \theta_i(t) \).

Let \( \bar{\tau} = \min\{\tau \in \{1, \ldots, t-1\} : \theta_i(\tau) + \epsilon \geq \theta_{i(t)}\} \).

If \( \bar{\tau} \) is not null, then \( i(t) = i(\bar{\tau}) \).

Essentially, at each time step \( t \), STS chooses to play an arm \( k \) that has already been played in the past, as long as the estimate of the expected reward from arm \( k \) is not lower than the estimate of the expected reward from an optimal arm (optimal at time \( t \) as per the TS) by \( \epsilon \) units. Thus, STS exploits more by re-using near-optimal satisficing arms rather than exploring unused arms. With per period regret as the performance metric...
(that captures the *time preference*), the simulation experiments reported in [29] show that STS can significantly outperform TS when the optimal action is costly to learn relative to satisfying near-optimal actions. We consider four simulation experiments that are same as those considered in [29] and compare the performance of TS-VHA-C1 with that of STS and TS in Fig. 8. For all the four experiments, we compute the per period regret over 500 time steps, averaged over 5000 runs. Fig. 6(a) considers a deterministic bandit with 250 arms with mean reward of each arm sampled independently from
As every arm, when played, returns the reward equal to its mean reward, it is referred to as a deterministic bandit. Performance of TS-VHA-C1-VA2 and TS-VHA-C1-VA3 is very close to that of STS. Fig. 6(b) corresponds to a bandit that differs from that of Fig. 6(a) as follows: Whenever an arm is played, the observed reward is a Bernoulli random variable with success probability equal to the mean reward. Note that, in Fig. 6(a) and Fig. 6(b), we consider $\epsilon = 0.05$ for the STS. It can be observed that TS-VHA-C1-VA2 and TS-VHA-C1-VA3 perform better than STS for time step values above (approximately) 150 and 50, respectively. Fig. 6(c) corresponds to a 250 armed Gaussian bandit with mean reward of each arm sampled independently from $\mathcal{N}(0, 1)$; When an arm is played, the realized reward is the sum of the arm’s mean reward and an independent sample from $\mathcal{N}(0, 1)$. Here, $\epsilon = 0.5$ for the STS. Finally, we consider linear Gaussian bandit with 250 arms in Fig. 6(d). The mean rewards are given by the vector $L\theta \in \mathbb{R}^{250 \times 1}$, where $\theta \in \mathbb{R}^{250 \times 1}$ is sampled from $\mathcal{N}(0, I)$ and $L \in \mathbb{R}^{250 \times 250}$ is a random matrix with each row drawn independently and uniformly from the unit sphere. While $\theta$ is unknown a priori, $L$ is known before hand. When an arm is played the observed reward is the sum of the mean reward and an independent sample from $\mathcal{N}(0, 2)$.

As can be observed from Fig. 6(c) and Fig. 6(d), TS-VHA-C1
Fig. 8: Cumulative regret comparison of TS-VHA-C3 with TS and TS-VHA-C3.

outperforms TS and STS for both independent Gaussian and Linear Gaussian bandits.

We think, the above simulation experiments only indicate that it might be interesting to investigate (and analyze) the TS-VHA-C1 from the aspect of time-sensitive learning.

D. Best Arm Identification

Next, we consider the fixed budget setting of the Best Arm Identification problem as discussed in [30]. The idea is to identify the best arm amongst all the bandit arms by playing them intelligently for a fixed number of time steps $t$. The metric used to compare algorithms is the probability of error in identifying the best arm after the fixed time step $t$.

TS can be utilized to solve this problem by designating the arm with the highest empirical mean after $t$ time steps as the best arm. However, TS performs poorly for this pure-exploration problem because of its high exploitative nature. Therefore, with the intention to increase the exploration in TS, we evaluate the applicability of TS-VHA-C2 in this scenario.

In Fig. 7, we consider the Bernoulli bandit as well as Gaussian bandit, each having two arms. The plots on the left and right have arms with mean rewards equal to $(0.5, 0.25)$ and $(0.51, 0.5)$, respectively. For both the scenarios, we observe that TS-VHA-C2 outperforms TS empirically.

E. Combiner C3

Finally, we evaluate the cumulative regret performance of Combiner C3, through simulations, for Bernoulli bandits and Gaussian bandits and compare its performance with TS and TS-VHA-C1.

Similar to section V-A and V-B, we first evaluate the performance of C3 on the randomized 20 arms case. As shown in Fig. 8(b) and Fig. 8(d), for both Gaussian and Bernoulli bandits, TS-VHA-C3 outperforms TS.

Next, we choose the same randomized scenario but with two arms in Fig. 8(a) and Fig. 8(c). In this case, C3 outperforms both TS and TS-VHA-C1 significantly. Interestingly, for the Gaussian bandits, TS-VHA-C1 performs inferior to TS, suggesting that, in some cases, increasing exploitation from the beginning does not help in optimizing the cumulative regret. But, dynamically adjusting the amount of exploitation over time by C3 provides superior performance. A mathematical analysis of the regret bound for C3 would help gaining more insight into it.

VI. Conclusion

We have proposed a general framework, Thompson Sampling with Virtual Helping Agents (TS-VHA), that combines samples drawn by the virtual agents to maneuver the exploration vs exploitation tradeoff in Thompson Sampling. Based on this framework, we developed two linear combiners (TS-VHA-C1 and TS-VHA-C2) and analysed theoretically their cumulative regret performance on Gaussian Bandits. Moreover, we showed their empirical efficacy on both Gaussian and Bernoulli bandits for multiple metrics: cumulative regret, best-arm identification and time-sensitive learning. We defer the analysis of the regret bounds on these metrics for our
future work. To exhibit the broad scope of the framework, we also put forth a nonlinear combiner TS-VHA-C3 that dynamically tunes the amount of exploration/exploitation and offers superior empirical performance. It would be interesting to experiment and devise more sophisticated combiners. TS-VHA can be applied wherever Thompson Sampling can be applied and we leave extending TS-VHA (along with designing combiners) for contextual bandits, non-stationary bandits and restless bandits for future work. Finally, exploring the usage of neural networks in developing combiners would be an exciting avenue for future work.

APPENDIX A
INEQUALITIES USED IN THE REGRET ANALYSIS

Inequality 1. (Chernoff-Hoeffding Bound) Let $X_1, \ldots, X_n$ be independent $0 \cdot 1$ r.v.s with $E[X_i] = p_i$ (not necessarily equal). Let $X = \frac{1}{n} \sum_i X_i$, $\mu = E[X] = \frac{1}{n} \sum_i p_i$. Then, for any $0 < \lambda < 1 - \mu$,
$$
Pr(X \geq \mu + \lambda) \leq e^{-n d(\mu + \lambda, \mu)},
$$
and for any $0 < \lambda < \mu$,
$$
Pr(X \geq \mu - \lambda) \leq e^{-n d(\mu - \lambda, \mu)},
$$
where $d(a, b) = a \ln \frac{a}{b} + (1 - a) \ln \frac{1 - a}{1 - b}$.

Inequality 2. (Chernoff-Hoeffding Bound) Let $X_1, \ldots, X_n$ be random variables with common range $[0, 1]$ and such that $E[X_i | X_1, \ldots, X_{i-1}] = \mu_i$. Let $S_n = \sum_{i=1}^n X_i$. Then, for all $a \geq 0$,
$$
Pr(S_n \geq a \mu + a) \leq e^{-a^2/2n},
$$
and
$$
Pr(S_n \leq a \mu - a) \leq e^{-a^2/2n}.
$$

The following inequalities can be derived for a Gaussian random variable from Formula 7.1.13 in [24].

Inequality 3. For a Gaussian distributed random variable $Z$ with mean $m$ and variance $\sigma^2$,
$$
Pr(Z > m + \sigma x) \geq \frac{x}{\sqrt{2\pi} (x^2 + 1)} e^{-x^2/2}.
$$

Inequality 4. For a Gaussian distributed random variable $Z$ with mean $m$ and variance $\sigma^2$, for any $z$,
$$
\frac{1}{4\sqrt{\pi}} e^{-z^2/2} < Pr(|Z - m| > \sigma z) \leq \frac{1}{2} e^{-z^2/2}.
$$

Inequality 5. Let $S_n = \sum_{i=1}^n \frac{1}{\tau_i}$. Then for $0 < p < 1$ from [31],
$$
S_n < 1 + \frac{(n + 1)^{1-p} - 1}{1 - p}.
$$

APPENDIX B
PROOF OF LEMMA 1

Recall that $\theta_i(t) \sim \mathcal{N}\left(\hat{\mu}_i(t), \frac{1}{\gamma k_i(t) + 1}\right)$, $x_i = \mu_i + \frac{\Delta_i}{\tau_i}$, $y_i = \mu_i - \frac{\Delta_i}{\tau_i}$ and $L_i(T) = \frac{2\ln T \Delta_i^2}{\gamma(y_i - \frac{\Delta_i}{\tau_i})^2}$. Given $\mathcal{F}_{\tau_k}$, let $\Theta_k$ denote a Gaussian random variable distributed as $\mathcal{N}\left(\hat{\mu}_1(\tau_k + 1), \frac{1}{\tau(k+1)}\right)$. For convenience, we denote $\hat{\mu}_1(\tau_k + 1)$ with $\hat{\mu}_1$ in the following. Let $G_k$ be the geometric random variable representing the number of consecutive independent trials until a sample of $\Theta_k$ becomes greater than $y_i$. Using $\Theta_k$ and Definition 6, we can write $p_i, \tau_k+1 = Pr(\Theta_k > y_i | \mathcal{F}_{\tau_k})$, and
$$
E\left[\frac{1 - p_i, \tau_k+1}{p_i, \tau_k+1}\right] = E[E[G_k | \mathcal{F}_{\tau_k}] = E[G_k].
$$

Therefore,
$$
\sum_{k=0}^{T-1} E\left[\frac{1 - p_i, \tau_k+1}{p_i, \tau_k+1}\right] = \frac{4L_i(T) - 1}{k=0} \sum_{k=0}^{T-1} E[G_k] + \frac{4L_i(T)}{k=0} \sum_{k=0}^{T-1} E[G_k]
$$
(19)

We will now bound Sum (19a), first term on the RHS of (19). Let $z = \sqrt{\ln r \beta}$, where $r \geq 1$ is an integer, $\beta \in [1, 2]$, and let $M_r$ denote the maximum of $r$ independent samples of $\Theta_k$.
$$
Pr(G_k < r) \geq Pr(M_r > y_i)
$$
$$
\geq Pr\left(M_r > \hat{\mu}_1 + \frac{z}{\sqrt{\gamma(k + 1)}} > y_i \right)
$$
$$
= E\left[Pr\left(M_r > y_i \mid \mathcal{F}_{\tau_k}\right)\right]
$$
$$
= E\left[I(\eta > y_i) Pr\left(M_r > y_i \mid \mathcal{F}_{\tau_k}\right)\right],
$$
(20)
where $\eta = \hat{\mu}_1 + \frac{z}{\sqrt{\gamma(k + 1)}}$. Since $\Theta_k \sim \mathcal{N}\left(\hat{\mu}_1, \frac{1}{\gamma(k + 1)}\right)$, using Inequality 3, we can write
$$
Pr\left(M_r > y_i \mid \mathcal{F}_{\tau_k} = \mathcal{F}_{\tau_k}\right) \geq 1 - \left(1 - \frac{1}{\sqrt{2\pi} \beta z^2 + 1} e^{-z^2/2}\right)^r
$$
$$
\geq 1 - e^{-\frac{z^2}{2r^2}}.
$$
(21)

Note that $\mathcal{F}_{\tau_k}$ is any realization of $\mathcal{F}_{\tau_k}$. As $\beta \in [1, 2]$, there exists a number $h(\beta) \in \mathbb{R}_{>0}$ such that $e^{-\frac{1}{2r^2}} \leq \frac{1}{r^2}$ for $r \geq h(\beta)$. Hence, for any $r \geq h(\beta)$ and any $\gamma > 0$,
$$
Pr\left(M_r > y_i \mid \mathcal{F}_{\tau_k} = \mathcal{F}_{\tau_k}\right) \geq 1 - \frac{1}{r^2}
$$
(22)

On substituting (22) in (20) we get, for any $r \geq h(\beta)$,
$$
Pr(G_k < r) \geq E\left[I(\eta > y_i) \left(1 - \frac{1}{r^2}\right)\right]
$$
$$
= \left(1 - \frac{1}{r^2}\right) Pr(\eta > y_i)
$$
(23)

We will now find a lower bound on $Pr(\eta > y_i)$.
$$
Pr(\eta > y_i) = Pr\left(\hat{\mu}_1 + \frac{z}{\sqrt{\gamma(k + 1)}} \geq \mu_1 - \frac{\Delta_i}{3}\right) = Pr\left(\hat{\mu}_1 + \frac{1}{k + 1} \geq \mu_1 - \frac{\Delta_i}{3} + \frac{z}{\sqrt{\gamma(k + 1)}} - \frac{1}{k + 1}\right)
$$
(24)

\frac{1}{\sqrt{(k+1)}} was added to $\hat{\mu}_1$ to account for the fact that $\hat{\mu}_1$ is not the average of the past $k$ observations, but it is the sum of the past $k$ observations divided by $(k + 1)$. Applying Inequality 2 to (24),
$$
Pr(\eta > y_i) \geq 1 - e^{-2\left(\Delta_i \sqrt{\frac{k+1}{3}} + \frac{z}{\sqrt{\gamma(k + 1)}} - \frac{1}{\sqrt{(k+1)}}\right)^2}.
$$
(25)
Substituting (25) back into (23), for any \( r \geq h(\beta) \),

\[
\Pr(G_k < r) \geq \left(1 - \frac{1}{r^2}\right) \left(1 - e^{-2\left(\frac{\Delta \gamma \sqrt{T}}{\sqrt{T} + \frac{3}{\sqrt{T} + 3}}\right)^2}\right)
\]

\[
\geq 1 - \frac{1}{r^2} - e^{-2\left(\frac{\Delta \gamma \sqrt{T}}{\sqrt{T} + \frac{3}{\sqrt{T} + 3}}\right)^2}.
\]  

(26)

This leads us to

\[
\sum_{k=0}^{4L_1(T)-1} \mathbb{E}[G_k] = \sum_{k=0}^{4L_1(T)-1} \sum_{r=0}^T \Pr(G_k \geq r) 
\]

\[
\leq \sum_{k=0}^{4L_1(T)-1} \sum_{r=0}^T \left(\frac{1}{r^2}\right) 
\]

\[
+ e^{-2\left(\frac{\Delta \gamma \sqrt{T}}{\sqrt{T} + \frac{3}{\sqrt{T} + 3}}\right)^2}
\]

(27)

First term on the RHS in (27) can be upper bounded as follows.

\[
\sum_{k=0}^{4L_1(T)-1} \left(\sum_{r=0}^T \frac{1}{r^2}\right) \leq \sum_{k=0}^{4L_1(T)-1} \left(h(\beta) + \sum_{r \geq h(\beta)} \frac{1}{r^2}\right) 
\]

\[
\leq \sum_{k=0}^{4L_1(T)-1} \left(h(\beta) + \zeta(2)\right) 
\]

\[
\leq (h(\beta) + \zeta(2))4L_1(T),
\]  

(28)

where \( \zeta \) is the Riemann zeta function. Next, we consider the second term on the RHS of (27) and use \( k' = k + 1 \) for convenience.

\[
\sum_{k'=1}^{4L_1(T)} \sum_{r=0}^T e^{-2\left(\frac{\Delta \gamma \sqrt{T}}{\sqrt{T} + \frac{3}{\sqrt{T} + 3}} - \frac{1}{T}\right)^2}
\]

\[
= \sum_{r=0}^T e^{-2\left(\frac{\Delta \gamma \sqrt{T}}{\sqrt{T} + \frac{3}{\sqrt{T} + 3}}\right)^2} e^{2\gamma(\sqrt{T} + \frac{3}{\sqrt{T} + 3})} e^{\frac{2\Delta \gamma}{3}}
\]

\[
\leq \sum_{r=0}^T e^{-2\left(\frac{\Delta \gamma \sqrt{T}}{\sqrt{T} + \frac{3}{\sqrt{T} + 3}}\right)^2} e^{2\gamma(\sqrt{T} + \frac{3}{\sqrt{T} + 3})} e^{\frac{4\Delta \gamma}{3}}
\]

\[
= \sum_{r=0}^T e^{-2\Delta \gamma \sqrt{T} + \frac{2\Delta \gamma \gamma}{\gamma - 1}} e^{-\frac{2\Delta \gamma \gamma}{2\gamma - 1}} e^{2\gamma(\sqrt{T} + \frac{3}{\sqrt{T} + 3})} e^{\frac{4\Delta \gamma}{3}}
\]

\[
(\alpha) \leq \sum_{r=0}^T e^{-2\Delta \gamma \sqrt{T} + \frac{2\Delta \gamma \gamma}{\gamma - 1}} e^{-\frac{2\Delta \gamma \gamma}{2\gamma - 1}} e^{2\gamma(\sqrt{T} + \frac{3}{\sqrt{T} + 3})} e^{\frac{4\Delta \gamma}{3}}
\]

\[
\leq \sum_{r=0}^T e^{-2\Delta \gamma \sqrt{T} + \frac{2\Delta \gamma \gamma}{\gamma - 1}} e^{2\gamma(\sqrt{T} + \frac{3}{\sqrt{T} + 3})} e^{\frac{4\Delta \gamma}{3}}
\]

\[
(\beta) \leq \sum_{r=0}^T e^{\frac{4\Delta \gamma}{3}} (\sqrt{T} + \frac{3}{\sqrt{T} + 3})
\]

(29)

\( (a) \) is due to the fact that

\[
\max\left(e^{-\frac{2\Delta \gamma \gamma}{2\gamma - 1}} (\sqrt{T} + \frac{3}{\sqrt{T} + 3})\right) = e^{2\gamma(\sqrt{T} + \frac{3}{\sqrt{T} + 3})}
\]

at \( k' = 1 \). In \( (b) \), \( c' = e^{\frac{4\Delta \gamma}{3}} / (e^{\frac{2\Delta \gamma \gamma}{2\gamma - 1}}) \). For any \( \epsilon > 0 \), there exists a number \( g(\epsilon) \) such that

\[
e^{\frac{4\Delta \gamma}{3}(1 - \frac{2\gamma}{\gamma - 1})} \leq \frac{1}{r^2 - \epsilon}
\]

for \( r \geq g(\epsilon) \). Hence, for \( \beta \in [1, 2] \), \( \gamma > 0 \), \( \epsilon > 0 \) and \( r \geq g(\epsilon) \),

\[
\sum_{k'=1}^{4L_1(T)} \sum_{r=0}^T e^{-2\left(\frac{\Delta \gamma \sqrt{T}}{\sqrt{T} + \frac{3}{\sqrt{T} + 3}} - \frac{1}{T}\right)^2} \leq \sum_{r=0}^T \frac{c'}{r^{2\gamma - \epsilon}}
\]

(30)

We will analyze (30) separately for \( \gamma \in (0, 4) \) and \( \gamma \geq 4 \).

For any value of \( \gamma \in (0, 4) \), we choose \( \beta \in [1, 2] \) such that \( \gamma < 2\beta \). Then, we select \( \epsilon > 0 \) to have \( \frac{2\beta}{\gamma} - \epsilon > 1 \). Thus, for \( \gamma \in (0, 4) \), (30) can be further simplified as,

\[
\sum_{k'=1}^{4L_1(T)} \sum_{r=0}^T e^{-2\left(\frac{\Delta \gamma \sqrt{T}}{\sqrt{T} + \frac{3}{\sqrt{T} + 3}} - \frac{1}{T}\right)^2} \leq \sum_{r=0}^T \frac{c'}{r^{2\gamma - \epsilon}}
\]

\[
\leq c' g(\epsilon) + \sum_{r \geq g(\epsilon)} \frac{c'}{r^{2\gamma - \epsilon}}
\]

\[
\leq c' \left( g(\epsilon) + \zeta\left(\frac{2\beta}{\gamma} - \epsilon\right)\right)
\]

(31)

Since \( \frac{2\beta}{\gamma} - \epsilon > 1 \) and \( \zeta \) is the Riemann zeta function, \( \zeta\left(\frac{2\beta}{\gamma} - \epsilon\right) \) is a finite number.

On the other hand, for \( \gamma \geq 4 \), \( \frac{2\beta}{\gamma} - \epsilon < 1 \) for any choice of \( \beta \) and \( \epsilon \). If we fix \( \beta \in [1, 2] \) and \( \epsilon > 0 \) such that \( \frac{2\beta}{\gamma} - \epsilon > 0 \) (30) results in,

\[
\sum_{k'=1}^{4L_1(T)} \sum_{r=0}^T \frac{\Delta \gamma \sqrt{T}}{\sqrt{T} + \frac{3}{\sqrt{T} + 3} - \frac{1}{T}}
\]

\[
\leq \sum_{r=0}^T \frac{c'}{r^{2\gamma - \epsilon}}
\]

\[
\leq c' \left( g(\epsilon) + \sum_{r \geq g(\epsilon)} \frac{c'}{r^{2\gamma - \epsilon}}\right)
\]

\[
\leq c' \left( g(\epsilon) + 1 + \frac{T^{1+\epsilon-\frac{2\beta}{\gamma} - 1}}{1+\epsilon-\frac{2\beta}{\gamma}}\right)
\]

(32)

The inequality (a) in (32) follows from Inequality 5. On substituting (31), (32) and (28) back into (27) gives us the bound for Sum (19a), the first term on the RHS of (19).

\[
\sum_{k=0}^{4L_1(T)-1} \mathbb{E}[G_k] \leq \begin{cases} 
H(\beta)L_1(T) + c' \left( g(\epsilon) + \zeta\left(\frac{2\beta}{\gamma} - \epsilon\right)\right) & \text{for } \gamma \in (0, 4), \\
H(\beta)L_1(T) + c' \left( g(\epsilon) + \frac{T^{1+\epsilon-\frac{2\beta}{\gamma} - 1}}{1+\epsilon-\frac{2\beta}{\gamma}}\right) & \text{for } \gamma \geq 4,
\end{cases}
\]

(33)

where, \( H(\beta) = 4(h(\beta) + \zeta(2)) \).

Next, we bound Sum (19b), second term on the RHS of (19) where the index of summation \( k \geq 4L_1(T) \). We will start by defining \( A_{t-1} \) as the event in which \( \mu_1(t) - \frac{d}{y_i} > y_i \) and use the notation \( \mathcal{F}_{t-1} \) to indicate random variable
conditioned on \( A_{k-1} \) being true, then,

\[
\mathbb{E} \left[ \frac{1}{p_{i, \tau_k+1}} \right] = \mathbb{E} \left[ \frac{1}{\Pr(\Theta_k > y_i | \mathcal{F}_{\tau_k})} \right] 
\leq \mathbb{E} \left[ \frac{1}{\Pr(\Theta_k > y_i | \mathcal{F}_{\tau_k} \cap A_{\tau_k})} \Pr(A_{\tau_k}) \right] 
\]

(34)

We now bound \( \Pr(\Theta_k > y_i | \mathcal{F}_{\tau_k} \cap A_{\tau_k}) \) and \( \Pr(A_{\tau_k}) \).

\[
\Pr(\Theta_k > y_i | \mathcal{F}_{\tau_k} \cap A_{\tau_k}) \geq \Pr(\Theta_k > \hat{\mu}_i - \frac{\Delta_i}{6} | \mathcal{F}_{\tau_k} \cap A_{\tau_k}) 
\geq 1 - e^{-\gamma(1+(k+1)\Delta_i^2)/72} 
\geq 1 - e^{-\gamma(4L_i(T))\Delta_i^2/72} 
\geq 1 - \frac{1}{T\Delta_i^2} 
\]

(35)

In the above, (a) follows from Inequality 2 with \( z = \sqrt{\gamma(k + 1)\Delta_i^2/6} \) and (b) is due to the fact that \( k \geq 4L_i(T) \).

Observe that for any \( t \geq \tau_k + 1 \), we have \( k_1(t) \geq k \geq 4L_i(T) \), and, using Inequality 2, we obtain,

\[
\Pr(A_{\tau_k}) = \Pr(\hat{\mu}_1(t) > \mu_1 - \frac{\Delta_i}{6}) \leq 1 - e^{-2\gamma k_1(t)\Delta_i^2/36} 
\geq 1 - \frac{1}{T\Delta_i^2} 
\]

(36)

Substituting (35) and (36) into (34), for \( k \geq 4L_i(T) \),

\[
\mathbb{E} \left[ \frac{1}{p_{i, \tau_k+1}} \right] - 1 \leq \frac{1}{\left(1 - \frac{1}{T\Delta_i^2}\right)^2} - 1 
\leq \frac{4}{T\Delta_i^2} 
\]

(37)

For any \( \gamma > 0 \), using (37), we get the following bound on Sum (19b).

\[
\sum_{t=\tau_k+1}^{T-1} \mathbb{E}[G_t] \leq \sum_{t=\tau_k}^{T-1} \frac{4}{T\Delta_i^2} 
\leq \frac{4}{\Delta_i^2} 
\]

(38)

Combining the results from (33), (38), (19) and (12) completes the proof of Lemma 1.
Kartik Anand Pant was born in Bhopal, Madhya Pradesh, India. He received his B.Tech. degree in electronics engineering from IIT (BHU) Varanasi, India, in 2018. He was a Scientist at Indian Space Research Organization, India, from 2018 to 2021. He is currently a Master’s student and a Graduate Research Assistant in Aeronautical and Astronautical Engineering at Purdue University. His research interests include Robotics and Control theory. He is a recipient of the Chintakindi Amba Rao Fellowship at Purdue University in 2021.

Amod Hegde was born in Sirsi, Karnataka, India, in 1996. He received the B.Tech. degree in electronics engineering from IIT (BHU) Varanasi, India, in 2018. He is currently a Master's student in computational social science at Stanford University. His research interests include optimization techniques/theory and machine learning. He was a recipient of the Director's Gold Medal at IIT BHU Varanasi, for outstanding all-round performance in the graduating class.

K. V. Srinivas was born in Vijayawada, India. He received the B.E. degree in electronics and communications engineering from the Andhra University College of Engineering, Vishakhapatnam, India, in June 1996, the M.Tech. degree from the Indian Institute of Technology, Kanpur, India, in 1998, and the Ph.D degree from the Indian Institute of Technology Madras, Chennai, India, in 2009, both in electrical engineering. He was a Postdoctoral Fellow at the Department of Electrical and Computer Engineering, University of Toronto, from March 2009 to October 2011. From January 2015 to December 2018, he was a Faculty member at the Indian Institute of Technology (BHU), Varanasi, India. His past industry experience includes working at the Indian Space Research Organisation, Samsung Electronics and Nokia Networks and Ericsson. In 2022 he joined Motorola Mobility/Lenovo as a researcher working on 5G NR standardization activities. His research interests include wireless communications, with emphasis on physical and MAC layer algorithms, and machine learning.