Batched Thompson Sampling

Cem Kalkanlı and Ayfer Özgür
Department of Electrical Engineering
Stanford University
{cemk, aozgur}@stanford.edu

Abstract

We introduce a novel anytime *batched* Thompson sampling policy for multi-armed bandits where the agent observes the rewards of her actions and adjusts her policy only at the end of a small number of batches. We show that this policy simultaneously achieves a problem dependent regret of order $O(\log(T))$ and a minimax regret of order $O(\sqrt{T \log(T)})$ while the number of batches can be bounded by $O(\log(T))$ independent of the problem instance over a time horizon T. We also prove that in expectation the instance dependent batch complexity of our policy is of order $O(\log \log(T))$. These results indicate that Thompson sampling performs competitively with recently proposed algorithms for the batched setting, which optimize the batch structure for a given time horizon T and prioritize exploration in the beginning of the experiment to eliminate suboptimal actions. Unlike these algorithms, the batched Thompson sampling algorithm we propose is an anytime policy, i.e. it operates without the knowledge of the time horizon T, and as such it is the only anytime algorithm that achieves optimal regret with $O(\log \log(T))$ expected batch complexity. This is achieved through a dynamic batching strategy, which uses the agents estimates to adaptively increase the batch duration.

1 Introduction

The multi-armed bandit problem models the scenario where an agent plays repeated actions and observes rewards associated with her actions. The agent aims to accumulate as much reward as possible, and consequently, she has to balance between playing arms that generated high rewards in the past, i.e. exploitation, and selecting under-explored arms that could potentially return better rewards, i.e. exploration. In the ideal scenario, the agent can adjust her policy once she receives feedback, e.g a reward, before the next action instance. However, many real-world applications limit the number of times the agent can interact with the system. For example, in medical applications [26], many patients are treated simultaneously in each trial, and the experimenter has to wait for the outcome of one trial before planing the next. In online marketing [25], there may be millions of responses per second, and as a result, it is not feasible for the advertiser to update her algorithm every time she receives feedback.

Recently, Perchet et al. [19] proposed to model this problem as the batched multi-armed bandits. Here, the experiment of duration T is split into a small number of batches and the agent does not receive any feedback regarding the rewards of its actions until the end of each batch. For the two-armed bandit problem, they proposed a general class of batched algorithms called explore-then-commit (ETC) policies, where the agent plays both arms the same number of times until the terminal batch and commits to the better performing arm in the last round unless the sample mean of one arm sufficiently dominates the other in earlier batches. They show that this algorithm achieves the optimal problem-dependent regret $O(\log(T))$ and the optimal minimax regret $O(\sqrt{T})$, matching the performance in the classical case where the agent receives instantaneous feedback about her actions, by using only $O(\log(T/\log(T)))$ and $O(\log \log(T))$ batches respectively. Their algorithm takes the time

horizon T and divides it into a fixed number of batches before the experiment where the specific batch structure is tuned to the target performance criteria, i.e. problem-dependent or minimax regret. Gao et al. [10] later generalized this result to the setting where the agent had more than two arms to choose from and she could adaptively adjust the batch sizes based on the past data. Their algorithm, called BaSE, is similar to the ETC algorithm in that in each batch the agent plays each of the remaining actions in a round robin fashion, and eliminates the underperforming arms at the end of each batch. They showed that this algorithm required $O(\log(T))$ and $O(\log\log(T))$ number of batches to achieve the optimal problem-dependent and the optimal minimax regret respectively, with batching strategies that were again tailored to the specific objective. More recently several other batched algorithms appeared in the context of asymptotic optimality [13], stochastic and adversarial bandits [9], and linear contextual bandits [11, 22, 20, 21], where the authors provided optimal algorithms in their respective settings. We note that there are also some earlier algorithms developed in the context of the classical bandit setting or bandits with switching cost that even-though not specially developed for the batched setting can be applied in a batched fashion [4, 6].

In this paper, we aim to study whether Thompson sampling, an algorithm that has been developed in 1933 [26] and since then successfully applied to a broad range of online optimization problems [7, 25], can achieve a similar performance in the batched setting. In the Thompson sampling algorithm, the agent chooses an action randomly according to its likelihood of being optimal, and after receiving feedback, i.e. observing rewards, updates its beliefs about the optimal action. The performance of Thompson sampling has been thoroughly analyzed in the literature [16, 17, 1, 2, 23, 24] and is known to achieve the optimal problem-dependent and minimax regret in the classical case. Our goal is to understand whether Thompson sampling can be combined with an adaptive batching strategy and maintain its regret performance when allowed to update its beliefs only at the end of a small number of batches. Note that the earlier algorithms developed specifically for the batched setting [19, 10] heavily prioritize exploration in the initial batches to eliminate the possibility that a suboptimal arm is played in the final exponentially larger batches, while Thompson sampling inherently balances between exploration and exploitation by randomly sampling actions according to their probability of being optimal. A first result in this direction has appeared in [14], which presents a batched Thompson sampling strategy called iPASE. However, [14] provides almost sure guarantees on the asymptotic regret which do not imply the guarantees on expected finite-time regret of interest in this paper and the previously mentioned literature.

Our main contribution in this paper is to show that Thompson sampling, combined with a novel adaptive batching scheme, achieves the optimal problem-dependent performance $O(\log(T))$ and at the same time a minimax regret of order $O(\sqrt{T\log(T)})$ by using $O(\log(T))$ batches independent of the problem instance. This performance is achieved simultaneously by a single strategy without the need to tune the batching structure according to the target criteria, i.e. problem-dependent or minimax regret. Moreover, unlike most of the previously mentioned batched algorithms where T is used both in action selection and the optimization of the batch structure, our strategy is an anytime strategy, i.e. it operates without the knowledge of the time horizon T. We note that policies designed for minimizing the problem-dependent regret can indeed be turned into anytime algorithms while retaining their $O(\log(T))$ regret and $O(\log(T))$ batch complexities with the help of the so called doubling trick in [5], but the same extension does not hold for minimax policies. This is because even with the best known doubling schemes (exponential or geometric), the minimax policies either suffer a regret significantly larger than $\sqrt{T}$ or have their batch complexity increase to $\Omega(\log(T))$ (exponential doubling leads to the first conclusion and geometric doubling leads to the second). This implies that our anytime Thompson sampling strategy matches the batch size of $O(\log(T))$ needed for these anytime extensions. In addition, we develop a problem-dependent bound on the expected number of batches used by our strategy which is $O(\log\log(T))$. This shows that while our strategy uses $O(\log(T))$ batches in the worst case, similar to previous algorithms, in a given instance of the problem with fixed reward distributions we only need $O(\log\log(T))$ batches on average. To the best of our knowledge, our scheme is the only strategy that can reduce the number of batches to $O(\log\log(T))$ without the knowledge of the time horizon T. This corresponds to a doubly exponential reduction in the interaction needed with the environment as compared to the classical case. We achieve this with a two-fold adaptive batching scheme that allows the batch sizes to increase doubly exponentially as the agent becomes more confident in its decisions. The scheme consists of two layers comprising of batches and multiple cycles inside each batch: the number of cycles per batch is increased exponentially, while at the same time each cycle becomes longer as the agent learns the environment and a suboptimal action is played less and less frequently. This notion of cycles is

2

key to allowing the batching scheme to dynamically adopt to each instance of the problem, which enables the $O(\log \log(T))$ guarantee on its expected problem-dependent batch complexity.

Finally, we would like to mention a concurrent and independent work by Karbasi et al. [15], also accepted to NeurIPS 2021. They consider the same problem and also develop a batch Thompson sampling strategy for the multi-armed bandits setting. While their results have some overlap with ours, there are also some differences. For example, they establish a $O(\log(T))$ worst-case guarantee on the batch complexity of their algorithm as we do in our paper, however their algorithm lacks a $O(\log \log(T))$ guarantee on its expected batch complexity. In contrast, they generalize their batch Thompson sampling strategy to the linear contextual setting, while we restrict our attention to multi-armed bandits.

2 Problem Formulation

2.1 Notations

We denote the natural logarithm as $\log(\cdot)$ while a logarithmic function of base $a > 1$ is $\log_a(\cdot)$. For the non-negative sequences of $\{a_n\}$ and $\{b_n\}$, $a_n = O(b_n)$ if and only if $\limsup_{n \to \infty} \frac{a_n}{b_n} < \infty$ and $a_n = \Omega(b_n)$ if and only if $b_n = O(a_n)$. We also denote by $Q(\cdot)$ the probability of a standard normal random variable X being bigger than a certain threshold x, i.e. $Q(x) = \mathbb{P}(X \geq x)$ for any $x \in \mathbb{R}$. Finally $\mathbb{1}(\cdot)$ is defined as the indicator function.

2.2 Batched Multi-Armed Bandit

We consider the batched multi-armed bandit setting. Here there are K arms, where each consecutive pull of the ith arm produces bounded i.i.d. rewards $\{Y_{i,t}\}_{t=1}^{\infty}$ such that

$$Y_{i,1} \in [0,1],$$
$$\mathbb{E}[Y_{i,1}] = \mu_i \in \mathbb{R}.$$

These mean rewards $\{\mu_i\}_{i=1}^{K}$ are assumed to be deterministic parameters unknown to the agent, whose goal is to accumulate as much reward as possible by repeatedly pulling these arms. Therefore, at each time instance t, the agent plays an arm $A_t \in \{1, 2, ..., K\}$ and receives the reward $Y_{A_t,t}$. Since she can only act causally and does not know $\{\mu_i\}_{i=1}^{K}$, she can only use the past observations, $\mathcal{H}_t = \{A_1, Y_{A_t,1}, ..., A_t, Y_{A_t,t}\}$ where $\mathcal{H}_0 = \emptyset$, to select the next action A_{t+1}.

In this paper, we study the batched version of this multi-armed bandit problem, where the agent has to play these arms in batches and can only incorporate the feedback from the system, i.e. her rewards, into her algorithm at the end of a batch. In other words, there are batch end points $0 = T_0 < T_1 < ...$, and the actions the agent plays in the jth batch $[T_{j-1} + 1, T_j]$ as well as the size of the batch itself, i.e. $T_j - T_{j-1}$, can depend only on the information present in $\mathcal{H}_{T_{j-1}}$ for any $j \in \mathbb{Z}^+$ and some external randomness that is independent of the system. Note that in this setting, the agent is allowed to adaptively choose the batch sizes depending on the past observations.

Finally, we let $\mu_1 > \mu_i$ for any $i \geq 2$. Given that the agent aims to maximize her cumulative reward, she would only play the first arm if she knew the hidden system parameters $\{\mu_i\}_{i=1}^{K}$. This observation naturally leads to the cumulative regret term, $R(T)$:

$$\mathbb{E}[R(T)] = \sum_{i=2}^{K} \Delta_i \, \mathbb{E}[N_i(T)] \tag{1}$$

where $\mu_1 - \mu_i = \Delta_i$ and

$$N_i(T) = \sum_{t=1}^{T} \mathbb{1}(A_t = i)$$

for $i \in \{1, 2, ..., K\}$.

3

3 Batched Thompson Sampling

In this section, we describe our Batched Thompson sampling strategy for the batched multi-armed bandit setting described in the previous section. This policy uses Gaussian priors in the spirit of [2] and each arm is sampled randomly according to its likelihood of being optimal under this prior and the observations from previous batches. We combine this strategy with an batching mechanism which relies on the notion of *cycles*. A cycle is defined as follows. The first cycle starts in the beginning of the experiment and ends when the agent selects an action different from the previous actions, i.e. it corresponds to the shortest time interval starting from the beginning of the experiment where two different actions are selected. Then the jth cycle for $j > 1$ is defined recursively as the time interval from the end of the $j - 1^{th}$ cycle to the first time step where the agent selects an action different from the first action in the cycle. In other words, in each cycle the agent plays exactly two different actions. Consider the following example. Assume that the first seven actions played by the agent are as follows: $A_1 = 1$, $A_2 = 1$, $A_3 = 2$, $A_4 = 1$, $A_5 = 3$, $A_6 = 2$, $A_7 = 2$. Then the first cycle is $[1, 3]$ because only at the third time step the agent selected an arm different from the earlier actions. Similarly, the second cycle is $[4, 5]$ where the agent played the first and the third actions. The third cycle that started on $t = 6$ has not ended yet because only a single action has been played so far. We use the concept of a cycle to adaptively decide on the batch size. At the beginning of the jth batch, the Thompson sampling agent checks the number of cycles in which each action i has been played since the beginning of the experiment, denoted $M_i(T_{j-1})$, and sets upper limits $U_{i,j} = \max\{1, \lceil \alpha \times M_i(T_{j-1}) \rceil\}$ for the cycle count of each action. Here $\alpha > 1$ is a batch growth factor to be chosen. Throughout the jth batch, the agent employs Thompson sampling, and at the end of each cycle checks whether or not the number of cycles in which a certain arm has been selected since the beginning of the experiment has reached its upper limit $U_{i,j}$ set for the current batch. The batch ends if there is one such action hitting its upper limit. After the jth batch, the agent observes the rewards of its actions and repeats the same process in the next batch. Note that as the algorithm proceeds the batch size increases due to two different reasons. First, the cycle count of each action is increased at the end of each batch. This alone leads to an exponential increase in the number of cycles per batch with base α. However, in addition to this, the length of each cycle becomes longer as the agent becomes more confident about its decisions and the probability of playing an optimal arm decreases. Note that the exponential increase in the number of cycles per batch is there even if the algorithm does not make progress in learning its environment, e.g. always plays arms with equal or near-equal probability due to a very noisy environment. However, the increase in the cycle duration is tightly coupled with the confidence of the agent and the cycle duration increases only if the agent becomes more confident. Therefore, this notion of cycles allows us to adopt the batch size much more dynamically to the current progress of the agent in a given instance of the problem.

We now introduce the following notation to denote the beginning and end of the kth cycle, $C_{b,k}$ and $C_{e,k}$ respectively:

$$C_{b,k} = \begin{cases} 1 & \text{if } k = 1 \\ C_{e,k-1} + 1 & \text{if } k > 1 \end{cases}$$

and

$$C_{e,k} = \min\{t \in \mathbb{Z}^+ | A_t \neq A_{t-1} \text{ and } t > C_{b,k}\}$$

for any positive integer k. As can be seen from these definitions, the interval $[C_{b,k}, C_{e,k}]$ describes the kth cycle. We also define $M_i(T)$ and $S_i(T)$ as follows

$$M_i(T) = \sum_{t=1}^{T} \mathbb{1}(A_t = i, t = C_{b,k} \text{ or } C_{e,k} \text{ for some } k)$$

and

$$S_i(T) = \sum_{t=1}^{T} \mathbb{1}(A_t = i, t = C_{b,k} \text{ or } C_{e,k} \text{ for some } k) Y_{i,t}.$$

Here $M_i(T)$ denotes the number of cycles in which the ith action has been selected, while $S_i(T)$ is the sum of rewards the agent received from playing the ith action at either the beginning or the end of a cycle over the duration T of the experiment. Note that whether the condition $\{t = C_{b,k} \text{ or } C_{e,k} \text{ for some } k\}$ is satisfied or not can be verified by checking the actions taken until the time step t, i.e. $\{A_j\}_{j=1}^{t}$. We also define $b(t) = \max\{j \in \mathbb{Z}_{\geq 0} | t - 1 \geq T_j\}$ as the index of the

last batch, and $B(t) = \min\{j \in \mathbb{Z}^+ | t \leq T_j\}$ as the batch index of the tth time step. We provide a pseudo-code for our Batched Thompson Sampling policy in Algorithm 1.

Algorithm 1: Batched Thompson Sampling

Input: Batch growth factor $\alpha > 1$, Gaussian variance σ^2
Initialization: $t = 1$, $M_i(0) = 0$, $S_i(0) = 0$, $U_{i,1} = 1$, $j = 1$, $T_0 = 0$.
while *Experiment Run* **do**

    Sample for Each Arm: $\theta_i(t) \sim \mathcal{N}(\frac{S_i(T_{j-1})}{1+M_i(T_{j-1})}, \frac{\sigma^2}{1+M_i(T_{j-1})})$
    Play an Arm: $A_t = \arg\max_i \theta_i(t)$
    Update the Pull Count: $M_i(t) \leftarrow M_i(t-1) + \mathbb{1}(A_t = i, t = C_{b,k} \text{ or } C_{e,k} \text{ for some } k)$
    if $t = C_{e,k}$ *for some k* **&** $M_i(t) = U_{i,j}$ *for some i* **then**
        End the Current Batch: $T_j = t$
        Receive the Rewards: $\{Y_{A_t,t}\}_{t=T_{j-1}+1}^{T_j}$
        Update the Cumulative Rewards for Each Arm:
        $S_i(T_j) = S_i(T_{j-1}) + \sum_{l=T_{j-1}+1}^{T_j} \mathbb{1}(A_l = i, l = C_{b,k} \text{ or } C_{e,k} \text{ for some } k)Y_{i,l}$
        Update the Upper Limites for the Cycle Counts: $U_{i,j+1} = \max\{1, \lceil \alpha \times M_i(T_j)\rceil\}$
        Update the Batch Index: $j \leftarrow j + 1$
    end
    Update the Time Index: $t \leftarrow t + 1$
end

Note that in Algorithm 1, the posterior distribution from which each arm is selected depends only on $\{M_i(T_{b(t)}), S_i(T_{b(t)})\}_{i=1}^K$, that is we only use the rewards from the first and last action selected in each cycle and ignore the rest of the observed rewards. This is to simplify the analysis in the following sections. However, we can also apply the algorithm by incorporating all the observed rewards, which in general can yield better performance while still maintaining the same batch structure. In that case, $\theta_i(t)$ for any t in the jth batch is drawn instead as

$$\theta_i(t) \sim \mathcal{N}\Big(\frac{\sum_{t=1}^{T_{j-1}} \mathbb{1}(A_t = i)Y_{i,t}}{1 + N_i(T_{j-1})}, \frac{\sigma^2}{1 + N_i(T_{j-1})}\Big).$$

4 Main Results

In this section, we state the main results of our paper. We start with the regret performance of Batched Thompson sampling.

Theorem 1. *Consider the batched multi-armed bandit setup described in Section 2. If $T \geq 2$ and the batch growth factor α satisfies $1 < \alpha \leq \frac{5\sigma^2}{4}$, then Batched Thompson sampling obeys the following inequalities*

$$\mathbb{E}[N_i(T)] \leq C_1 \alpha \sigma^2 \frac{\log(T)}{\Delta_i^2} \tag{2}$$

for any $i \geq 2$, which lead to

$$\mathbb{E}[R(T)] \leq C_1 \alpha \sigma^2 \sum_{i=2}^{K} \frac{\log(T)}{\Delta_i}, \tag{3}$$

and

$$\mathbb{E}[R(T)] \leq C_2 \sigma \sqrt{\alpha K T \log(T)} \tag{4}$$

where $C_1, C_2 \geq 1$ are absolute constants independent of the system parameters.

We provide the proof of this theorem for the special case of $K = 2$, $\alpha = 2$, $\sigma^2 = 1$ in Section 6.1, and defer the proof of the general version to the supplementary materials.

Theorem 1 states that Batched Thompson sampling achieves $O(\log(T) \sum_{i=2}^{K} \Delta_i^{-1})$ problem-dependent regret and $O(\sqrt{KT \log(T)})$ minimax regret, which match the asymptotic lower bound of $\Omega(\log(T))$ [18] and the minimax lower bound of $\Omega(\sqrt{KT})$ [3] up to a $\sqrt{\log(T)}$ term respectively.

We next compare our bounds with the results on classical Thompson sampling [2]. As described in the previous section, we use Gaussian priors for Thompson sampling following the work of Agrawal and Goyal [2]. This is one of the two priors considered in that work for Thompson sampling: beta priors and Gaussian priors. For Thompson sampling with Beta priors, Agrawal and Goyal [2] provides two different bounds on the expected regret:

$$\mathbb{E}[R(T)] \leq (1 + \epsilon) \sum_{i=2}^{K} \frac{\log(T)}{d(\mu_i, \mu_1)} \Delta_i + O(\frac{K}{\epsilon^2}), \tag{5}$$

for any $\epsilon \in (0, 1)$, and

$$\mathbb{E}[R(T)] \leq O(\sqrt{KT \log(T)}), \tag{6}$$

where $d(\mu_i, \mu_1) = \mu_i \log(\frac{\mu_i}{\mu_1}) + (1 - \mu_i) \log(\frac{1-\mu_i}{1-\mu_1})$. In addition, they show that with Gaussian priors the expected regret of the classical Thompson sampling is bounded by $O(\sqrt{KT \log(K)})$ if $T \geq K$. Considering that $d(\mu_i, \mu_1) \geq 2\Delta_i^2$ by Pinsker's inequality, (5) provides a tighter performance guarantee than (3) in terms of the dependence on the reward distributions, but we note that the minimax performance of Batched Thompson sampling, (4), matches the performance of classical Thompson sampling when the agent receives instantaneous feedback about rewards and can update its policy after each action. These results show that Batched Thompson sampling, apart from the dependence on the reward distributions in the problem dependent bound, matches the regret performance in the classical case.

We note that the regret bounds in the theorem depend on the batch growth factor α. The regret increases linearly as α grows bigger; this is not surprising since bigger batch sizes mean fewer updates for Batched Thompson sampling.

We now present the batch complexity results for our algorithm.

Theorem 2. *Consider the batched multi-armed bandit setup described in Section 2. If $T \geq 2$ and $1 < \alpha \leq \frac{5\sigma^2}{4}$, then the batch complexity $B(T)$ of Batched Thompson sampling satisfies the following:*

$$B(T) \leq 1 + K + K \log_\alpha(1 + \frac{T}{K}), \tag{7}$$

$$\mathbb{E}[B(T)] \leq 1 + K + \log_\alpha(1 + C\alpha\sigma^2 \sum_{i=2}^{K} \frac{\log(T)}{\Delta_i^2})) + \sum_{i=2}^{K} \log_\alpha(1 + C\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2}), \text{ and} \tag{8}$$

$$\mathbb{E}[B(T)] \leq 1 + 2C\alpha\sigma^2 \sum_{i=2}^{K} \frac{\log(T)}{\Delta_i^2}, \tag{9}$$

where C is an absolute constant independent of the system parameters.

Theorem 2 states three different batch complexity guarantees; the first is a deterministic guarantee on the number of batches while the last two bound the number of batches in expectation. If we consider (7), we see that Batched Thompson sampling uses at most $O(\log(T))$ many batches regardless of the reward distributions. This result and Theorem 1 indicate that Batched Thompson sampling matches the regret performance and the batch complexities of optimal problem-dependent batched algorithms developed in [4, 10, 9, 15], which also achieve $O(K \log(T))$ problem-dependent regret with $O(\log(T))$ batch complexity. However, compared with the other optimal algorithms, we show that we can further reduce the expected batch complexity down to $O(K \log \log(T))$ in (8). This is because our batching strategy uses the information it gathers about the system (through the notion of cycles) to adaptively decide on the sizes of the batches while most prior batched algorithms use a static batch structure. We note that Algorithm 1 of Esfandiari et al. [9] does use an adaptive batching strategy however that strategy appears to be geared towards obtaining a tighter regret bound rather than reducing the batch complexity.

We note that the bounds on the number of batches in (7) and (8) diverge to infinity as $\alpha \downarrow 1$. The bound in (9) on the other hand decreases with α and can be relevant when α is chosen very small. We also note that if $\alpha < 1 + \frac{1}{T}$ for a fixed T, then Batched Thompson sampling will only allow one cycle per batch throughout the experiment of duration T and the notion of a batch will coincide with the

notion of a cycle. In this extremal case with one cycle per batch, (9) shows that Batched Thompson sampling will have $O(\log(T))$ batch complexity, while still satisfying the regret bounds stated in Theorem 1. The $O(K \log \log(T))$ bound on the expected batch complexity in (8) is enabled by the fact that for larger α we allow for exponentially more cycles in each batch. As the algorithm proceeds and becomes more confident about the system, choosing a suboptimal arm becomes less likely and as a result the cycle durations become inherently larger. At the same time, the algorithm allows for exponentially more cycles in each batch. This double batching strategy in our algorithm, via cycles and batches is key to obtain $O(K \log \log(T))$ guarantee in (8) with an anytime strategy. Note that the best previously available guarantee on batch complexity for an anytime batching strategy is $O(\log(T))$.

The proof of these theorems are provided in Section 6. The main technical novelty in the proof comes from the fact that the notion of cycles leads to a random duration for each batch without any deterministic upper bound on the number of times each arm is played in the batch. This makes it more difficult to control the regret accumulated during a given batch. We note that without the notion of cycles (e.g. when a cycle corresponds to a pull of an arm) the exponential increase in the number of arms per batch is relatively easy to deal with, as in this case we have a deterministic upper bound on the number of times each arm can be played in a given batch.

5 Experiments

In this section, we provide some experimental results on the performance of Batched Thompson sampling where we do not skip samples, i.e. the variant mentioned at the end of Section 3, for different values of α, $\{1.00001, 1.25, 1.5, 2\}$, and how they perform against normal Thompson sampling under different reward distributions and action counts when time horizon $T = 5 \times 10^4$ and the sampling variance $\sigma^2 = 1$. We mainly consider four setups: Bernoulli rewards when $K = 2$, Figure 1 (a); Bernoulli rewards when $K = 5$, Figure 1 (b); Gaussian rewards when $K = 2$, Figure 1 (c); Gaussian rewards when $K = 5$, Figure 1 (d). Finally each figure is the result of an experiment averaged over 10^4 repeats and the average number of batches used throughout the experiment is rounded up to the nearest integer and reported in the parenthesis to the right of the algorithm names on the figures. The figures show that our batching strategy matches the performance of classical Thompson sampling by using roughly 100 batches over a time horizon of $T = 5 \times 10^4$.

As can be seen from Figure 1, Batched Thompson sampling achieves almost the same empirical performance as the normal Thompson sampling when we set α small enough so that there is only one cycle per batch, i.e. $\alpha = 1.00001$. We also observe that this Batched Thompson sampling version $\alpha = 1.00001$, can have a batch count as small as 15. However when α is very small, the problem independent guarantees in (7) and (8) become very loose and the number of batches can vary more with the reward distributions. This can be partly observed in the figures: for $\alpha = 1.00001$, there is larger variation in the average batch complexity across different reward distributions though in all cases the average numbe of batches remain very small. Increasing α leads to a more stable batch complexity behavior, at the cost of a small multiplicative regret factor; we observe that the batch count almost remains constant for $\alpha = 2$ across different reward distributions. The source codes of the experiment can be found in `https://github.com/incsmi/BatchedThompsonSampling.git`.

6 Technical Analysis

In this section we provide technical proofs for our results. We start with the proof of Theorem 1 for a special case and at the end prove Theorem 2.

6.1 Proof of Theorem 1 when $K = 2$, $\alpha = 2$, and $\sigma^2 = 1$

We first introduce $N_{2,j}(t)$ as the number of times the second arm is pulled if Batched Thompson sampling is employed for t many round with the past knowledge of $\mathcal{H}_{T_{j-1}}$. In this case, $N_{2,j}(T_j - T_{j-1}) = N_2(T_j) - N_2(T_{j-1})$. We know that in the first T rounds, there can be no more than T many batches, and each batch can not last longer than T rounds. As a result, we have the following bound
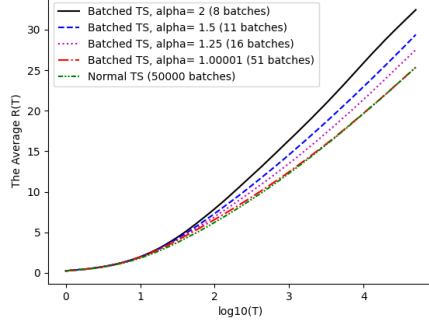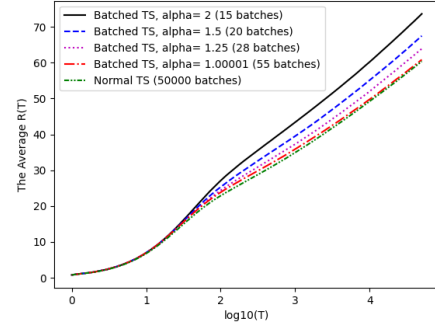
(a) $Y_{1,t} \sim \text{Bern}(0.75), Y_{2,t} \sim \text{Bern}(0.25)$
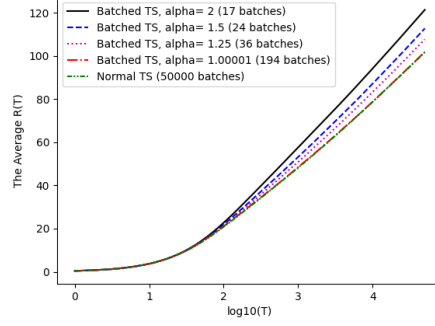
(b) $Y_{1,t} \sim \text{Bern}(0.75), Y_{i,t} \sim \text{Bern}(0.25)\ 2 \le i \le 5$
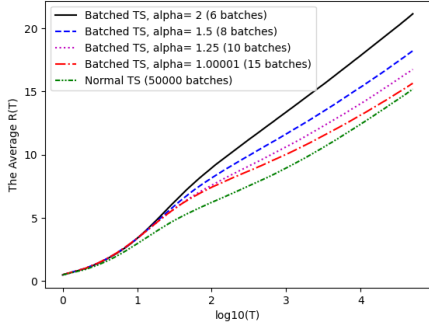
(c) $Y_{1,t} \sim \mathcal{N}(1,1), Y_{2,t} \sim \mathcal{N}(0,1)$

(d) $Y_{1,t} \sim \mathcal{N}(1,1), Y_{i,t} \sim \mathcal{N}(0,1)\ 2 \le i \le 5$

Figure 1: Empirical Regret Performance of Batched and normal Thompson sampling

on $N_2(T)$:

$$N_2(T) \le \sum_{j=1}^{T} \min(N_{2,j}(T_j - T_{j-1}), N_{2,j}(T)).$$

Now we first analyze the expected number of times the second is pulled in the first cycle of the jth batch. It is easy to see the time the first arm is selected in this cycle is an upper bound on the number of times the second arm is picked. This observation follows from the fact that if the first action is selected in the first round, then the second is selected only once in the current cycle, if not the time the first arm is selected becomes one more than the number of times the second arm is picked. As a result, conditioned on the past $\mathcal{H}_{T_{j-1}}$, the expected number of times the second arm is picked in a single cycle is upper bounded by $\frac{1}{\mathbb{P}(A_{T_{j-1}+1}=1|H_{T_{j-1}})}$. Also note that since there are only two actions, each cycle will contain both of them, and in the first j batches, there will be 2^{j-1} many cycles by the construction of Algorithm 1. This observation means that there are no more than 2^{j-1} many identically distributed such cycles in the jth batch, and we have the following bound for any j:

$$\mathbb{E}[N_{2,j}(T_j - T_{j-1})] \le 2^{j-1} \mathbb{E}[\frac{1}{\mathbb{P}(A_{T_{j-1}+1} = 1|H_{T_{j-1}})}] \le C2^{j-1} \tag{10}$$

where $C \ge 2$ is a constant independent of j. The last inequality in (10) follows from the following Lemma 3 and the fact that in the first batch $\mathbb{P}(A_{T_{j-1}+1} = 1|H_{T_{j-1}}) = \mathbb{P}(A_1 = 1) = \frac{1}{2}$.

Lemma 3. *If $K = 2$, $\alpha = 2$, and $\sigma^2 = 1$, then any $j \ge 2$:*

$$\mathbb{E}\left[\frac{1}{\mathbb{P}(A_{T_{j-1}+1} = 1|H_{T_{j-1}})}\right] \le C, \tag{11}$$

$$\mathbb{P}(A_{T_{j-1}+1} = 2) \le \exp(-\frac{2^{j-4}}{3}\Delta_2^2), \tag{12}$$

8

where $C \geq 2$ is a constant independent of j.

In addition, we also have:

$$\mathbb{E}[N_{2,j}(T)] = T\,\mathbb{P}(A_{T_{j-1}+1} = 2) \leq T\exp(-\frac{2^{j-4}}{3}\Delta_2^2) \tag{13}$$

for $j \geq 2$ by the same lemma. The overall analysis shows that for any positive integer k, we have:

$$\mathbb{E}[N_2(T)] \leq \mathbb{E}[\sum_{j=1}^{T}\min(N_{2,j}(T_j - T_{j-1}), N_{2,j}(T))] \leq \sum_{j=1}^{k}\mathbb{E}[N_{2,j}(T_j - T_{j-1})] + \sum_{j=k+1}^{T}\mathbb{E}[N_{2,j}(T)]$$

$$\leq C(2^k - 1) + T\sum_{j=k+1}^{T}\exp(-\frac{2^{j-4}}{3}\Delta_2^2) \tag{14}$$

where the last step follows from (10) and (13). Let k be the smallest positive integer such that $\frac{2^k}{3} \geq 8\frac{\log(T)}{\Delta_2^2}$. Then we have

$$T\sum_{j=k+1}^{T}\exp(-\frac{2^{j-4}}{3}\Delta_2^2) \leq T\sum_{i=0}^{\infty}\exp(-2^i\frac{2^{k-3}}{3}\Delta_2^2) \leq T\sum_{i=0}^{\infty}\exp(-2^i\log(T))$$

$$= \sum_{i=0}^{\infty}\frac{1}{T^{2^i - 1}} \leq \sum_{i=0}^{\infty}\frac{1}{T^i} \leq 2$$

where the last inequality follows from the assumption that $T \geq 2$. This analysis bounds the last summation term in (14). To bound $C(2^k - 1)$, note that k is the smallest positive integer bigger than $\log_2(24\frac{\log(T)}{\Delta_2^2})$. Since $\Delta_2^2 \leq 1$ and $T \geq 2$, we know $24\frac{\log(T)}{\Delta_2^2} \geq 1$. This analysis shows that $k \leq \log_2(24\frac{\log(T)}{\Delta_2^2})+1 = \log_2(48\frac{\log(T)}{\Delta_2^2})$. As a result, $C(2^k - 1) \leq 48C\frac{\log(T)}{\Delta_2^2}$. The overall analysis shows that $\mathbb{E}[R(T)] = \Delta_2\mathbb{E}[N_2(T)] \leq 48C\frac{\log(T)}{\Delta_2} + 2\Delta_2$ by (14). This finishes the proof of (3). Finally (4) is proven by (3) if $\Delta_2 > \sqrt{\frac{\log(T)}{T}}$. If not, then $\mathbb{E}[R(T)] \leq T\Delta_2 \leq \sqrt{T\log(T)}$, and this proves (4).

6.2   Proof of Theorem 2

Let us consider the case where the agent has already employed the Batched Thompson sampling, Algorithm 1, for T many steps and denote $i_j \in \{1, 2, ..., B(T) - 1\}$ for $j \in \{1, 2, ..., k_i\}$ as the indices where $M_i(T_{i_j}) = U_{i,i_j}$. Since each batch end point T_j has to satisfy the condition $M_l(T_j) = U_{l,j}$ for some l, we have

$$B(T) - 1 \leq \sum_{i=1}^{K}k_i. \tag{15}$$

By the definition of $U_{i,j}$, we know that $M_i(T_{i_1}) = 1$. In addition, note that there may be batches in between i_{j-1}^{th} and i_j^{th} ones and the agent may have picked the ith arm while the condition $M_i(T_j) = U_{i,j}$ is not satisfied. These observations lead to $\max\{\alpha M_i(T_{i_{j-1}}), M_i(T_{i_{j-1}}) + 1\} \leq M_i(T_{i_j})$ for $j \geq 2$. The overall analysis shows that if $k_i \geq 1$, then $\alpha^{k_i-1} \leq M_i(T_{i_{k_i}}) \leq M_i(T_{B(T)-1})$ due to the fact that $T_{i_{k_i}} \leq T_{B(T)-1}$, which leads to $k_i \leq 1 + \log_\alpha(1 + M_i(T_{B(T)-1}))$ for any $k_i \geq 0$. In addition, we also have the following trivial bound $k_i \leq M_i(T_{i_{k_i}}) \leq M_i(T_{B(T)-1})$. As a result of these inequalities and (15), we have

$$B(T) \leq 1 + K + \sum_{i=1}^{K}\log_\alpha(1 + M_i(T_{B(T)-1})) \tag{16}$$

and

$$B(T) \leq 1 + \sum_{i=1}^{K}M_i(T_{B(T)-1}). \tag{17}$$

9

First of all, since $\log(\cdot)$ is a concave function, Jensen's inequality and (16) lead to

$$B(T) \leq 1 + K + K \log_\alpha (1 + \frac{1}{K} \sum_{i=1}^{K} M_i(T_{B(T)-1})). \tag{18}$$

Considering that each cycle has to contain at least two action steps, there can be no more than $\frac{T}{2}$ many cycles in the first $B(T) - 1$ batches. In addition, each cycle can only be recorded once by two different actions. This observation leads to $\sum_{i=1}^{K} M_i(T_{B(T)-1}) \leq T$, which proves (7) by (18). To prove (8), we first note that $M_1(T_{B(T)-1}) \leq \sum_{i=2}^{K} M_i(T_{B(T)-1})$ because each cycle of the first arm has to be accompanied by another arm. Since $M_i(T_{B(T)-1}) \leq N_i(T)$ for any i, we have $B(T) \leq 1 + K + \log_\alpha (1 + \sum_{i=2}^{K} N_i(T)) + \sum_{i=2}^{K} \log_\alpha (1 + N_i(T))$ by (16), which shows that

$$\mathbb{E}[B(T)] \leq \mathbb{E}[1 + K + \log_\alpha (1 + \sum_{i=2}^{K} N_i(T)) + \sum_{i=2}^{K} \log_\alpha (1 + N_i(T))]$$

$$\leq 1 + K + \log_\alpha (1 + \sum_{i=2}^{K} \mathbb{E}[N_i(T)]) + \sum_{i=2}^{K} \log_\alpha (1 + \mathbb{E}[N_i(T)]).$$

The last inequality follows from Jensen's inequality. This leads to (8) by the fact that $\mathbb{E}[N_i(T)] \leq C\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2}$ from (2). Finally, the previous analysis also implies that $B(T) \leq 1 + 2\sum_{i=2}^{K} N_i(T)$ by (17). This inequality and (2) prove (9).

7 Conclusion

We proposed an anytime Batched Thompson sampling algorithm and proved that it achieves the optimal problem-dependent and minimax regret with only $O(\log(T))$ instance-independent batch complexity, matching the state-of-the-art anytime batched algorithms. More interestingly, we showed that in a given instance of the problem our algorithm requires only $O(\log\log(T))$ batches on average, which corresponds to a doubly exponential decrease in the interaction needed with the environment as compared to the classical case. To the best of our knowledge, previous anytime algorithms only satisfy a worst-case $O(\log(T))$ guarantee on the batch complexity. Finally, simulations show that Batched Thompson sampling performs empirically close to classical Thompson sampling by using drastically fewer batches.

8 Acknowledgements

References

[1] Marc Abeille, Alessandro Lazaric, et al. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

[2] Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

[3] Jean-Yves Audibert, Sébastien Bubeck, et al. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

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

[5] Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

[6] Nicolo Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems*, pages 1160–1168, 2013.

[7] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[8] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[9] Hossein Esfandiari, Amin Karbasi, Abbas Mehrabian, and Vahab Mirrokni. Regret bounds for batched bandits. *arXiv preprint arXiv:1910.04959*, 2019.

[10] Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. In *Advances in Neural Information Processing Systems*, pages 503–513, 2019.

[11] Yanjun Han, Zhengqing Zhou, Zhengyuan Zhou, Jose Blanchet, Peter W Glynn, and Yinyu Ye. Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*, 2020.

[12] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[13] Tianyuan Jin, Pan Xu, Xiaokui Xiao, and Quanquan Gu. Double explore-then-commit: Asymptotic optimality and beyond. *arXiv preprint arXiv:2002.09174*, 2020.

[14] Cem Kalkanlı and Ayfer Özgür. Asymptotic performance of thompson sampling in the batched multi-armed bandits. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 539–544. IEEE, 2021.

[15] Amin Karbasi, Vahab Mirrokni, and Mohammad Shadravan. Parallelizing thompson sampling. *arXiv preprint arXiv:2106.01420*, 2021.

[16] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

[17] Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in neural information processing systems*, pages 1448–1456, 2013.

[18] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[19] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, Erik Snowberg, et al. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681, 2016.

[20] Zhimei Ren and Zhengyuan Zhou. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *arXiv preprint arXiv:2008.11918*, 2020.

[21] Zhimei Ren, Zhengyuan Zhou, and Jayant R Kalagnanam. Batched learning in generalized linear contextual bandits with general decision sets. *IEEE Control Systems Letters*, 2020.

[22] Yufei Ruan, Jiaqi Yang, and Yuan Zhou. Linear bandits with limited adaptivity and learning distributional optimal design. *arXiv preprint arXiv:2007.01980*, 2020.

[23] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[24] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

[25] Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

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

[27] Roman Vershynin. High-dimensional probability, 2019.

A Outline

The appendix is organized as follows.

1. Section B states technical tools necessary for our proofs.
2. Section C provides the proof of Lemma 3, which is stated in Section 6.1.
3. In Section D, we present couple propositions and lemmas in preparation for the proof of Theorem 1 in the general case of K, α, and σ^2.
4. The full proof of Theorem 1 is finally given in Section E.
5. In Section F, we provide additional experiments in which we compare different variants of Batched Thompson sampling: the one using all the observations and the one skipping samples of multiple instances from the same arm.

B Technical Tools

B.1 Bounded Random Variable Moment-generating Function Bound

Let X be a bounded zero mean random variable such that $a \leq X \leq b$ and $|a|, |b| < \infty$. Hoeffding [12] showed that

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2 \frac{(b-a)^2}{8}) \tag{19}$$

for any real number λ.

B.2 Gaussian Tail Bounds

Proposition 2.1.2 of [27] shows that

$$\left(\frac{1}{\delta} - \frac{1}{\delta^3}\right) \frac{\exp(-\frac{\delta^2}{2})}{\sqrt{2\pi}} \leq Q(\delta) \leq \frac{1}{\delta} \frac{\exp(-\frac{\delta^2}{2})}{\sqrt{2\pi}}, \tag{20}$$

if $\delta > 0$. Since exponential functions decay faster than power functions, there exists δ_0 such that if $\delta \geq \delta_0$, then

$$\exp(-3\delta^2/4) \leq Q(\delta)$$

which leads to

$$Q^{-1}(1/x) \geq \sqrt{\frac{4}{3}\log(x)} \tag{21}$$

if $x \geq x_0$ for some $x_0 \geq 2$, where $Q^{-1}(\cdot)$ is the inverse function of $Q(\cdot)$. Note that the last inequality follows from setting $\delta = \sqrt{\frac{4}{3}\log(x)}$ and the fact that $Q(\cdot)$ is decreasing.

B.3 Expectation of Non-negative Random Variables

Let X be a non-negative random variable, i.e. $X \geq 0$, then

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)dx \tag{22}$$

by Lemma 2.2.13 of [8].

C Proof of Lemma 3

First of all,

$$\mathbb{E}\left[\frac{1}{\mathbb{P}(A_{T_{j-1}+1} = 1|H_{T_{j-1}})}\right] = \int_0^\infty \mathbb{P}\left(\frac{1}{\mathbb{P}(A_{T_{j-1}+1} = 1|H_{T_{j-1}})} > x\right)dx$$

$$\leq 2 + \int_2^\infty \mathbb{P}(\mathbb{P}(A_{T_{j-1}+1} = 1|H_{T_{j-1}}) \leq \frac{1}{x})dx \tag{23}$$

by (22). Conditioned on $\mathcal{H}_{T_{j-1}}$, $\theta_i(T_{j-1}+1)$ is distributed as $\mathcal{N}(\frac{S_i(T_{j-1})}{1+M_i(T_{j-1})}, \frac{1}{1+M_i(T_{j-1})})$. However when $K = 2$ and $\alpha = 2$, we know that $M_i(T_{j-1}) = 2^{j-2}$. This overall analysis leads to

$$
\begin{aligned}
\mathbb{P}(A_{T_{j-1}+1} = 1 | H_{T_{j-1}}) &= \mathbb{P}\left(\mathcal{N}\left(\frac{S_1(T_{j-1})}{1+2^{j-2}}, \frac{1}{1+2^{j-2}}\right) \geq \mathcal{N}\left(\frac{S_2(T_{j-1})}{1+2^{j-2}}, \frac{1}{1+2^{j-2}}\right) | H_{T_{j-1}}\right) \\
&= \mathbb{P}\left(\mathcal{N}\left(\frac{S_1(T_{j-1}) - S_2(T_{j-1})}{1+2^{j-2}}, \frac{2}{1+2^{j-2}}\right) \geq 0 | H_{T_{j-1}}\right) \\
&= 1 - \mathbb{P}\left(\mathcal{N}\left(0, \frac{2}{1+2^{j-2}}\right) \geq \frac{S_1(T_{j-1}) - S_2(T_{j-1})}{1+2^{j-2}} | H_{T_{j-1}}\right) \\
&= 1 - Q\left(\frac{S_1(T_{j-1}) - S_2(T_{j-1})}{\sqrt{2+2^{j-1}}}\right)
\end{aligned}
$$

where the last equality follows from the definition of the function $Q(\cdot)$. Combining the last analysis with (23) shows that

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{\mathbb{P}(A_{T_{j-1}+1} = 1 | H_{T_{j-1}})}\right] &\leq 2 + \int_2^\infty \mathbb{P}\left(1 - Q\left(\frac{S_1(T_{j-1}) - S_2(T_{j-1})}{\sqrt{2+2^{j-1}}}\right) \leq \frac{1}{x}\right) dx \\
&= 2 + \int_2^\infty \mathbb{P}\left(Q\left(\frac{S_1(T_{j-1}) - S_2(T_{j-1})}{\sqrt{2+2^{j-1}}}\right) \geq 1 - \frac{1}{x}\right) dx \\
&= 2 + \int_2^\infty \mathbb{P}\left(\frac{S_2(T_{j-1}) - S_1(T_{j-1})}{\sqrt{2+2^{j-1}}} \geq -Q^{-1}\left(1 - \frac{1}{x}\right)\right) dx \quad (24) \\
&= 2 + \int_2^\infty \mathbb{P}\left(\frac{S_2(T_{j-1}) - S_1(T_{j-1})}{\sqrt{2+2^{j-1}}} \geq Q^{-1}\left(\frac{1}{x}\right)\right) dx \quad (25)
\end{aligned}
$$

where (24) follows from the fact that $Q(\cdot)$ is a decreasing function, and (25) is the result of the symmetric nature of the normal distribution.

Let λ be any real number. Here we are going to use induction hypothesis. We know that $\mathbb{E}[\exp(\lambda(S_2(T_1) - S_1(T_1) + \Delta_2))] \leq \exp(\lambda^2/4)$ by (19) and the fact that $S_1(T_1)$ and $S_2(T_1)$ are independent bounded random variables. Now assume that $\mathbb{E}[\exp(\lambda(S_2(T_j) - S_1(T_j) + 2^{j-1}\Delta_2))] \leq \exp(2^{j-3}\lambda^2)$ for some $j \geq 1$. However we know that $\mathcal{H}_{T_j}$, $S_1(T_{j+1}) - S_1(T_j)$, and $S_2(T_{j+1}) - S_2(T_j)$ are mutually independent. This is because regardless of what the agent observes in the first j batches, i.e. $\mathcal{H}_{T_j}$, she is going to record rewards from both arms only 2^{j-1} numbers times in the $j + 1^{th}$ batch and it does not matter at which time indices these are recorded since all the future rewards from any arm are i.i.d. as well. This leads to

$$
\begin{aligned}
&\mathbb{E}[\exp(\lambda(S_2(T_{j+1}) - S_1(T_{j+1}) + 2^j \Delta_2))] \\
&= \mathbb{E}[\exp(\lambda(S_2(T_j) - S_1(T_j) + 2^{j-1}\Delta_2)) \\
&\qquad\qquad \mathbb{E}[\exp(\lambda(S_2(T_{j+1}) - S_2(T_j) - S_1(T_j) + S_1(T_{j+1}) + 2^{j-1}\Delta_2)) | \mathcal{H}_{T_j}]] \quad (26)
\end{aligned}
$$

However from the earlier analysis and the fact that the $j + 1^{th}$ batch contains 2^{j-1} recorded rewards from each arm, we have

$$
\begin{aligned}
&\mathbb{E}[\exp(\lambda(S_2(T_{j+1}) - S_2(T_j) - S_1(T_j) + S_1(T_{j+1}) + 2^{j-1}\Delta_2)) | \mathcal{H}_{T_j}] \\
&= \mathbb{E}[\exp(\lambda(S_2(T_{j+1}) - S_2(T_j) - 2^{j-1}\mu_2))] \mathbb{E}[\exp(-\lambda(S_1(T_{j+1}) - S_1(T_j) - 2^{j-1}\mu_1))] \\
&\leq \exp(2^{j-3}\lambda^2) \quad (27)
\end{aligned}
$$

where the last inequality follows from (19) and the fact that the first l rewards and $l + 1^{th}$ reward from the same arm are independent since the rewards are i.i.d. Note that each arm will be picked infinitely often since probability selecting any arm in any batch will be almost surely positive due to using Gaussian distribution to select arms. Finally, (26) and (27), along with the induction step, shows that

$$
\mathbb{E}[\exp(\lambda(S_2(T_{j-1}) - S_1(T_{j-1}) + 2^{j-2}\Delta_2))] \leq \exp(2^{j-4}\lambda^2) \quad (28)
$$

for any $j \geq 2$. This result leads to the following bound for any $\lambda \geq 0$ and $x \geq 2$:

$$
\begin{aligned}
\mathbb{P}\left(\frac{S_2(T_{j-1}) - S_1(T_{j-1})}{\sqrt{2+2^{j-1}}} \geq Q^{-1}\left(\frac{1}{x}\right)\right) &\leq \mathbb{P}\left(\frac{S_2(T_{j-1}) - S_1(T_{j-1}) + 2^{j-2}\Delta_2}{\sqrt{2+2^{j-1}}} \geq Q^{-1}\left(\frac{1}{x}\right)\right) \\
&\leq \exp(\lambda^2/8 - \lambda Q^{-1}(1/x))
\end{aligned}
$$

13

where the last inequality follows from the Chernoff bound. Since $Q^{-1}(1/x) \geq 0$ when $x \geq 2$, setting $\lambda = 4Q^{-1}(1/x)$ shows that

$$\mathbb{P}\left(\frac{S_2(T_{j-1}) - S_1(T_{j-1})}{\sqrt{2 + 2^{j-1}}} \geq Q^{-1}\left(\frac{1}{x}\right)\right) \leq \exp(-2(Q^{-1}(1/x))^2)$$

for any $x \geq 2$. Finally, by (21)

$$\mathbb{P}\left(\frac{S_2(T_{j-1}) - S_1(T_{j-1})}{\sqrt{2 + 2^{j-1}}} \geq Q^{-1}\left(\frac{1}{x}\right)\right) \leq \frac{1}{x^{8/3}} \tag{29}$$

if $x \geq x_0$. Putting (29) back into (25) leads to:

$$\mathbb{E}\left[\frac{1}{\mathbb{P}(A_{T_{j-1}+1} = 1 | H_{T_{j-1}})}\right] \leq x_0 + \int_{x_0}^{\infty} \frac{1}{x^{8/3}} dx \leq x_0 + 1,$$

which proves (11) since x_0 is independent of any system parameter.

We now prove (12). Similar to the earlier analysis, we can describe the probability of selecting the second arm as the sample from the second arm being bigger than the first arm's:

$$\mathbb{P}(A_{T_{j-1}+1} = 2) = \mathbb{P}\left(\mathcal{N}\left(\frac{S_2(T_{j-1}) - S_1(T_{j-1})}{1 + 2^{j-2}}, \frac{2}{1 + 2^{j-2}}\right) \geq 0\right)$$
$$= \mathbb{P}\left(\mathcal{N}\left(\frac{S_2(T_{j-1}) - S_1(T_{j-1})}{1 + 2^{j-2}} + \frac{2^{j-2}}{1 + 2^{j-2}}\Delta_2, \frac{2}{1 + 2^{j-2}}\right) \geq \frac{2^{j-2}}{1 + 2^{j-2}}\Delta_2\right)$$

In view of (28), we know that $S_2(T_{j-1}) - S_1(T_{j-1}) + 2^{j-2}\Delta_2$ is sub Gaussian with variance proxy 2^{j-3}. As a result, $\mathcal{N}\left(\frac{S_2(T_{j-1}) - S_1(T_{j-1})}{1 + 2^{j-2}} + \frac{2^{j-2}}{1 + 2^{j-2}}\Delta_2, \frac{2}{1 + 2^{j-2}}\right)$ has the following variance proxy:

$$\frac{2^{j-3}}{(1 + 2^{j-2})^2} + \frac{2}{1 + 2^{j-2}} = \frac{2^{j-3} + 2(1 + 2^{j-2})}{(1 + 2^{j-2})^2}.$$

This observation and Chernoff bound, which states that $\mathbb{P}(X \geq x) \leq \exp(-x^2/(2\sigma^2))$ if $x \geq 0$ and X is sub Gaussian with variance proxy σ^2, lead to:

$$\mathbb{P}(A_{T_{j-1}+1} = 2) \leq \exp\left(-\frac{(1 + 2^{j-2})^2}{2^{j-2} + 4(1 + 2^{j-2})} \frac{(2^{j-2})^2}{(1 + 2^{j-2})^2}\Delta_2^2\right)$$
$$= \exp\left(-\frac{2^{2j-4}}{2^{j-2} + 4(1 + 2^{j-2})}\Delta_2^2\right)$$
$$\leq \exp\left(-\frac{2^{2j-4}}{3 \times 2^j}\Delta_2^2\right)$$
$$= \exp\left(-\frac{2^{j-4}}{3}\Delta_2^2\right)$$

which finishes the proof of (12).

D   Results Related to Theorem 1

D.1   Martingale Lemma

In this part, we present a key martingale lemma.

Lemma 4. *Let $\mathcal{F}_t = \{Y_{A_1,1}, Y_{A_1,1}, ..., Y_{A_{T_b(t)}, T_{b(t)}}, A_1, A_2, ..., A_t\}$, then $X_t = \exp(\lambda(S_i(T_{b(t)}) - \mu_i M_i(T_{b(t)})) - \frac{\lambda^2}{8}(1 + M_i(T_{b(t)})))$ is a non-negative supermartingale adapted to $\mathcal{F}_t$ for any real λ and $i \in \{1, 2, .., K\}$. Finally for any t we have*

$$\mathbb{E}[X_t] \leq 1, \tag{30}$$

and in particular any stopping time $\tau \leq \infty$ for $\{\mathcal{F}_t\}$ satisfies the following inequality

$$\mathbb{E}[X_\tau] \leq 1, \tag{31}$$

where $\lim_{t \to \infty} X_t = X_\infty$.

Proof. First of all, it is clear that X_t's are integrable, since the rewards $\{Y_{A_t,t}\}$ are bounded. That means the only thing we need to prove is that $\{X_t\}$ is a supermartingale sequence, i.e. the following inequality

$$\mathbb{E}[X_{t+1}|\mathcal{F}_t] \leq X_t \tag{32}$$

almost surely for any positive integer t. By the definition $b(t)$ we know that $b(t)$ and $b(t+1)$ are functions of $\{A_1, A_2, ..., A_t\} \subseteq \mathcal{F}_t$. Note that the batch end points are decided by the actions taken. In addition, on $\{b(t) = b(t+1)\}$ X_t is equal to X_{t+1}. These observations lead to

$$\mathbb{E}[X_{t+1}|\mathcal{F}_t] = \mathbb{E}[\exp(\lambda(S_i(T_{b(t+1)}) - \mu_i M_i(T_{b(t+1)})) - \frac{\lambda^2}{8}(1 + M_i(T_{b(t+1)})))|\mathcal{F}_t]$$

$$= \mathbb{1}(b(t) = b(t+1))X_t$$

$$+ \mathbb{E}[\mathbb{1}(b(t) \neq b(t+1))\exp(\sum_{j=T_{b(t)}+1}^{T_{b(t+1)}} \mathbb{1}(A_j = i, j = C_{b,k} \text{ or } C_{e,k} \text{ for some } k)(\lambda Y_{i,j} - \lambda\mu_i - \frac{\lambda^2}{8}))|\mathcal{F}_t]X_t \tag{33}$$

where the last equality follows from the definitions of $S_i(T_{b(t)})$ and $M_i(T_{b(t)})$. Note that $b(t) \neq b(t+1)$ if and only if t is a batch end point, which leads to

$$\mathbb{1}(b(t) \neq b(t+1)) = \sum_{l=0}^{t-1} \mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l). \tag{34}$$

We first prove that $\mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l)$ and $\{Y_{i,j}\}_{j=l+1}^{t}$ are independent. To that end, it is enough to show that

$$\mathbb{P}(T_{b(t+1)} = t, T_{b(t)} = l, (Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S}) = \mathbb{P}(T_{b(t+1)} = t, T_{b(t)} = l)\mathbb{P}((Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S})$$

for any Borel set $\mathcal{S}$ of $\mathbb{R}^{t-l}$. Note that $T_{b(t)} = l$ and $T_{b(t+1)} = t$ if and only if l is a batch end point, call this event E_1, and the smallest batch end point strictly bigger than l is t, call this event E_2. This means $\{T_{b(t)} = l, T_{b(t+1)} = t\} = E_1 \cap E_2$. Then we know $\mathbb{P}((Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S}, E_1) = \mathbb{P}((Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S})\mathbb{P}(E_1)$ due to the fact that whether or not l is a batch end point depends on the past actions $\{A_j\}_{j=1}^{l}$, which are independent of the future rewards from the ith arm $\{Y_{i,j}\}_{j=l+1}^{t}$. In addition, conditioned on the fact that there is a batch end point at l, i.e. the event E_1, different values for $\{Y_{i,j}\}_{j=l+1}^{t}$ won't change the probability of E_2 happening. This is because the agent can not use the information present in $\{Y_{i,j}\}_{j=l+1}^{t}$ unless the current that started at $l+1$ ends. As a result, we have $\mathbb{P}(E_2|E_1, (Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S}) = \mathbb{P}(E_2|E_1)$. The overall analysis leads to

$$\mathbb{P}(T_{b(t+1)} = t, T_{b(t)} = l, (Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S}) = \mathbb{P}(E_1, E_2, (Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S})$$

$$= \mathbb{P}((Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S})\mathbb{P}(E_1|(Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S})\mathbb{P}(E_2|E_1, (Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S})$$

$$= \mathbb{P}((Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S})\mathbb{P}(E_1)\mathbb{P}(E_2|E_1)$$

$$= \mathbb{P}((Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S})\mathbb{P}(E_1, E_2)$$

$$= \mathbb{P}((Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S})\mathbb{P}(T_{b(t+1)} = t, T_{b(t)} = l), \tag{35}$$

which finishes the proof of the fact that $\mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l)$ and $\{Y_{i,j}\}_{j=l+1}^{t}$ are independent. Similar to the earlier analysis, conditioned on $\{T_{b(t+1)} = t, T_{b(t)} = l\}$, $\{A_1, Y_{A_1,1}, A_2, Y_{A_2,2}, ..., A_{T_{b(t)}}, Y_{A_{T_{b(t)}},T_{b(t)}}\}$ and $\{Y_{i,j}\}_{j=l+1}^{t}$ are independent, because the future rewards from the ith arm can not affect the past observations, i.e. actions and rewards. In addition, conditioned on $\{T_{b(t+1)} = t, T_{b(t)} = l\}$, we know that the actions $\{A_j\}_{j=T_{b(t)}+1}^{t}$ are sampled according to the information present in $\{A_1, Y_{A_1,1}, A_2, Y_{A_2,2}, ..., A_{T_{b(t)}}, Y_{A_{T_{b(t)}},T_{b(t)}}\}$. As a result, conditioned on $\{T_{b(t+1)} = t, T_{b(t)} = l\}$, $\mathcal{F}_t$ and $\{Y_{i,j}\}_{j=l+1}^{t}$ are independent. This overall analysis shows that for any Borel set $\mathcal{S}$ of $\mathbb{R}^{t-l}$ and any element $\mathcal{G}$ of the sigma algebra generated by $\mathcal{F}_t$ we have

$$\mathbb{P}(T_{b(t+1)} = t, T_{b(t)} = l, (Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S}, \mathcal{G})$$

$$= \mathbb{P}(T_{b(t+1)} = t, T_{b(t)} = l, (Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S})\mathbb{P}(\mathcal{G}|T_{b(t+1)} = t, T_{b(t)} = l, (Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S})$$

$$= \mathbb{P}((Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S})\mathbb{P}(T_{b(t+1)} = t, T_{b(t)} = l)\mathbb{P}(\mathcal{G}|T_{b(t+1)} = t, T_{b(t)} = l) \tag{36}$$

$$= \mathbb{P}((Y_{i,l+1}, ..., Y_{i,t}) \in \mathcal{S})\mathbb{P}(T_{b(t+1)} = t, T_{b(t)} = l, \mathcal{G}), \tag{37}$$

where (36) follows from (35) and the fact that conditioned on $\{T_{b(t+1)} = t, T_{b(t)} = l\}$, $\mathcal{F}_t$ and $\{Y_{i,j}\}_{j=l+1}^{t}$ are independent. Since $\mathcal{S}$ is arbitrary, $\{Y_{i,j}\}_{j=l+1}^{t}$ and $\mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l, \mathcal{G})$ are independent.

Now we go back to (34), and note that $\mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l)$ can be written as a sum of the terms of the following form

$$\mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l) \prod_{n=l+1}^{t} (\mathbb{1}(a_n = -1) + a_n \times \mathbb{1}(A_n = i, n = C_{b,k} \text{ or } C_{e,k} \text{ for some } k)),$$
(38)

where $a_n \in \{-1, 1\}$. Note that the terms of this form are indicator functions with disjoint domains. Then we have for any element $\hat{\mathcal{G}}$ of the sigma algebra generated by $\mathcal{F}_t$

$$\mathbb{E}[\mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l)(\prod_{n=l+1}^{t} (\mathbb{1}(a_n = -1) + a_n \times \mathbb{1}(A_n = i, n = C_{b,k} \text{ or } C_{e,k} \text{ for some } k))) \mathbb{1}(\hat{\mathcal{G}})$$

$$\exp(\sum_{j=T_{b(t)}+1}^{T_{b(t+1)}} \mathbb{1}(A_j = i, j = C_{b,k} \text{ or } C_{e,k} \text{ for some } k)(\lambda Y_{i,j} - \lambda \mu_i - \frac{\lambda^2}{8}))]$$

$$= \mathbb{E}[\mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l)(\prod_{n=l+1}^{t} (\mathbb{1}(a_n = -1) + a_n \times \mathbb{1}(A_n = i, n = C_{b,k} \text{ or } C_{e,k} \text{ for some } k))) \mathbb{1}(\hat{\mathcal{G}})$$

$$\exp(\sum_{j=l+1}^{t} \mathbb{1}(a_j = 1)(\lambda Y_{i,j} - \lambda \mu_i - \frac{\lambda^2}{8}))] \qquad (39)$$

Note that a_n's are deterministic variables and by the earlier analysis, i.e. (37), we know that $\mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l)(\prod_{n=l+1}^{t}(\mathbb{1}(a_n = -1) + a_n \times \mathbb{1}(A_n = i, n = C_{b,k} \text{ or } C_{e,k} \text{ for some } k))) \mathbb{1}(\hat{\mathcal{G}})$ and $\{Y_{i,j}\}_{j=l+1}^{t}$ are independent. Then by the fact that $\mathbb{E}[\exp(\lambda Y_{i,j} - \lambda \mu_i - \frac{\lambda^2}{8})] \leq 1$ due to (19), we have

$$\mathbb{E}[\mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l)(\prod_{n=l+1}^{t} (\mathbb{1}(a_n = -1) + a_n \times \mathbb{1}(A_n = i, n = C_{b,k} \text{ or } C_{e,k} \text{ for some } k))) \mathbb{1}(\hat{\mathcal{G}})$$

$$\exp(\sum_{j=T_{b(t)}+1}^{T_{b(t+1)}} \mathbb{1}(A_j = i, j = C_{b,k} \text{ or } C_{e,k} \text{ for some } k)(\lambda Y_{i,j} - \lambda \mu_i - \frac{\lambda^2}{8}))]$$

$$\leq \mathbb{E}[\mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l)(\prod_{n=l+1}^{t} (\mathbb{1}(a_n = -1) + a_n \times \mathbb{1}(A_n = i, n = C_{b,k} \text{ or } C_{e,k} \text{ for some } k))) \mathbb{1}(\hat{\mathcal{G}})].$$

Since $\hat{\mathcal{G}}$ is arbitrary, the previous inequality shows that

$$\mathbb{E}[\mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l)(\prod_{n=l+1}^{t} (\mathbb{1}(a_n = -1) + a_n \times \mathbb{1}(A_n = i, n = C_{b,k} \text{ or } C_{e,k} \text{ for some } k)))$$

$$\exp(\sum_{j=T_{b(t)}+1}^{T_{b(t+1)}} \mathbb{1}(A_j = i, j = C_{b,k} \text{ or } C_{e,k} \text{ for some } k)(\lambda Y_{i,j} - \lambda \mu_i - \frac{\lambda^2}{8})) | \mathcal{F}_t]$$

$$\leq \mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l)(\prod_{n=l+1}^{t} (\mathbb{1}(a_n = -1) + a_n \times \mathbb{1}(A_n = i, n = C_{b,k} \text{ or } C_{e,k} \text{ for some } k)))$$

almost surely, which leads to

$$\mathbb{E}[\mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l) \exp(\sum_{j=T_{b(t)}+1}^{T_{b(t+1)}} \mathbb{1}(A_j = i, j = C_{b,k} \text{ or } C_{e,k} \text{ for some } k)(\lambda Y_{i,j} - \lambda \mu_i - \frac{\lambda^2}{8})) | \mathcal{F}_t]$$

$$\leq \mathbb{1}(T_{b(t+1)} = t, T_{b(t)} = l) \qquad (40)$$

16

almost surely by the observation in (38). Finally combining (34) and (40) proves that

$$\mathbb{E}[\mathbb{1}(b(t) \neq b(t+1)) \exp(\sum_{j=T_{b(t)}+1}^{T_{b(t+1)}} \mathbb{1}(A_j = i, j = C_{b,k} \text{ or } C_{e,k} \text{ for some } k)(\lambda Y_{i,j} - \lambda \mu_i - \frac{\lambda^2}{8}))| \mathcal{F}_t]$$

$$\leq \mathbb{1}(b(t) \neq b(t+1)) \tag{41}$$

almost surely. This inequality and (33) lead to (32). We have showed that X_t is a supermartingale sequence.

Finally, we prove (30) and (31). Firstly, note that $T_{b(1)} = T_0 = 0$, $S_i(0) = 0$, and $M_i(0) = 0$, which lead to $\mathbb{E}[X_t] \leq \mathbb{E}[X_1] = \exp(-\lambda^2/8) \leq 1$ for any t by the properties of supermartingales. Coupling this fact with the following theorem finishes the proof:

Theorem 4.8.4 of [8]. *If X_t is a non-negative supermartingale and $\tau \leq \infty$ is a stopping time, then $\mathbb{E}[X_\tau] \leq \mathbb{E}[X_1]$ where $\lim_{t \to \infty} X_t$ exists and $X_\infty = \lim_{t \to \infty} X_t$.*

$\square$

D.2 Estimation Error Bound

In this section, we provide a proposition stating that if a certain arms is selected in sufficiently many cycles, then sample θ_i corresponding to that arm has to be close to the true mean with high probability.

Proposition 5. *Let $T \geq 2$, then for any positive integer t and $i \in \{1, 2, ..., K\}$, we have*

$$\mathbb{P}\left(\theta_1(t) \leq \frac{\mu_1 + \mu_i}{2}, M_1(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right) \leq \frac{2}{T} \tag{42}$$

and

$$\mathbb{P}\left(\theta_i(t) > \frac{\mu_1 + \mu_i}{2}, M_i(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right) \leq \frac{2}{T}. \tag{43}$$

Proof. We first prove (42). Here we have

$$\mathbb{P}\left(\theta_1(t) \leq \frac{\mu_1 + \mu_i}{2}, M_1(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right)$$

$$\leq \mathbb{P}\left(\theta_1(t) \leq \frac{\mu_1 + \mu_i}{2}, \frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} \geq \frac{3\mu_1 + \mu_i}{4}, M_1(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right)$$

$$+ \mathbb{P}\left(\frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} < \frac{3\mu_1 + \mu_i}{4}, M_1(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right) \tag{44}$$

We know that conditioned on $\mathcal{H}_{t-1}$, $\theta_1(t)$ is distributed as $\mathcal{N}(\frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})}, \frac{\sigma^2}{1 + M_1(T_{b(t)})})$. This fact leads to

$$\mathbb{P}(\theta_1(t) \leq \frac{\mu_1 + \mu_i}{2} | \mathcal{H}_{t-1}) = Q\left(\frac{\sqrt{1 + M_1(T_{b(t)})}}{\sigma}\left(\frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} - \frac{\mu_1 + \mu_i}{2}\right)\right).$$

Then we have

$$\mathbb{P}\left(\theta_1(t) \leq \frac{\mu_1 + \mu_i}{2}, \frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} \geq \frac{3\mu_1 + \mu_i}{4}, M_1(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right)$$

$$= \mathbb{E}[\mathbb{P}(\theta_1(t) \leq \frac{\mu_1 + \mu_i}{2} | \mathcal{H}_{t-1}) \mathbb{1}\left(\frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} \geq \frac{3\mu_1 + \mu_i}{4}, M_1(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right)]$$

$$= \mathbb{E}[Q\left(\frac{\sqrt{1 + M_1(T_{b(t)})}}{\sigma}\left(\frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} - \frac{\mu_1 + \mu_i}{2}\right)\right)$$

$$\mathbb{1}\left(\frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} \geq \frac{3\mu_1 + \mu_i}{4}, M_1(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right)] \tag{45}$$

17

where the first equality follows from the fact that $S_1(T_{b(t)})$ and $M_1(T_{b(t)})$ are measurable with respect to $\mathcal{H}_{t-1}$. Since

$$\frac{\sqrt{1 + M_1(T_{b(t)})}}{\sigma}\left(\frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} - \frac{\mu_1 + \mu_i}{2}\right) \geq \sqrt{2\log(T)}$$

on $\{\frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} \geq \frac{3\mu_1 + \mu_i}{4}, M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2}\}$, (45) leads to

$$\mathbb{P}\left(\theta_1(t) \leq \frac{\mu_1 + \mu_i}{2}, \frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} \geq \frac{3\mu_1 + \mu_i}{4}, M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2}\right) \leq Q(\sqrt{2\log(T)}).$$

However, we know that $Q(x) \leq \exp(-x^2/2)$ for $x \geq 1$ by (20), which results in

$$\mathbb{P}\left(\theta_1(t) \leq \frac{\mu_1 + \mu_i}{2}, \frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} \geq \frac{3\mu_1 + \mu_i}{4}, M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2}\right) \leq \frac{1}{T}. \qquad (46)$$

We now bound the second term on the right-hand side of (44). Note that

$$\mathbb{P}\left(\frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} < \frac{3\mu_1 + \mu_i}{4}, M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2}\right)$$

$$= \mathbb{P}\left(\frac{S_1(T_{b(t)}) - \mu_1 M_1(T_{b(t)})}{1 + M_1(T_{b(t)})} < \frac{3\mu_1 + \mu_i}{4} - \mu_1\frac{M_1(T_{b(t)})}{1 + M_1(T_{b(t)})}, M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2}\right)$$

$$= \mathbb{P}\left(\frac{S_1(T_{b(t)}) - \mu_1 M_1(T_{b(t)})}{1 + M_1(T_{b(t)})} < -\frac{\Delta_i}{4} + \mu_1\frac{1}{1 + M_1(T_{b(t)})}, M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2}\right). \qquad (47)$$

We know that $\mu_1 \leq 1$ and

$$\frac{\mu_1}{1 + M_1(T_{b(t)})} \leq \frac{\Delta_i^2}{32\sigma^2\log(T)} \leq \frac{\Delta_i}{8}$$

if $M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2}$. Note that the last inequality follows from the fact that $0 \leq \Delta_i \leq 1$ and $32\sigma^2\log(T) \geq 8$ for $T \geq 2$ and $\sigma^2 \geq 1$. This analysis and (47) indicate that

$$\mathbb{P}\left(\frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} < \frac{3\mu_1 + \mu_i}{4}, M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2}\right)$$

$$\leq \mathbb{P}\left(\frac{S_1(T_{b(t)}) - \mu_1 M_1(T_{b(t)})}{1 + M_1(T_{b(t)})} < -\frac{\Delta_i}{8}, M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2}\right). \qquad (48)$$

Then by Lemma 4 with $\lambda = -\frac{\Delta_i}{2}$

$$1 \geq \mathbb{E}[\exp(-\frac{\Delta_i}{2}(S_1(T_{b(t)}) - \mu_1 M_1(T_{b(t)})) - \frac{\Delta_i^2}{32}(1 + M_1(T_{b(t)})))$$

$$\mathbb{1}(\frac{S_1(T_{b(t)}) - \mu_1 M_1(T_{b(t)})}{1 + M_1(T_{b(t)})} < -\frac{\Delta_i}{8}, M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2})]$$

$$\geq \mathbb{E}[\exp(\frac{\Delta_i^2}{32}(1 + M_1(T_{b(t)}))) \mathbb{1}(\frac{S_1(T_{b(t)}) - \mu_1 M_1(T_{b(t)})}{1 + M_1(T_{b(t)})} < -\frac{\Delta_i}{8}, M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2})]$$

$$\tag{49}$$

$$\geq T\,\mathbb{P}\left(\frac{S_1(T_{b(t)}) - \mu_1 M_1(T_{b(t)})}{1 + M_1(T_{b(t)})} < -\frac{\Delta_i}{8}, M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2}\right) \tag{50}$$

where (49) from the condition set inside the indicator function, i.e. $\frac{S_1(T_{b(t)}) - \mu_1 M_1(T_{b(t)})}{1 + M_1(T_{b(t)})} < -\frac{\Delta_i}{8}$.
Similarly, the condition $M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2}$ and $\sigma^2 \geq 1$ lead to (50). Then combining (48) and (50) results in

$$\mathbb{P}\left(\frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} < \frac{3\mu_1 + \mu_i}{4}, M_1(T_{b(t)}) \geq 32\sigma^2\frac{\log(T)}{\Delta_i^2}\right) \leq \frac{1}{T}$$

18

which, along with (44) and (46), proves (42). We now prove (43). However its proof is almost the same as the proof of (42). Similarly we have

$$\mathbb{P}\left(\theta_i(t) > \frac{\mu_1 + \mu_i}{2}, M_i(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right)$$

$$\leq \mathbb{P}\left(\theta_i(t) > \frac{\mu_1 + \mu_i}{2}, \frac{S_i(T_{b(t)})}{1 + M_i(T_{b(t)})} \leq \frac{\mu_1 + 3\mu_i}{4}, M_i(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right)$$

$$+ \mathbb{P}\left(\frac{S_i(T_{b(t)})}{1 + M_i(T_{b(t)})} > \frac{\mu_1 + 3\mu_i}{4}, M_i(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right) \tag{51}$$

Here conditioned on $\mathcal{H}_{t-1}$, $\theta_i(t)$ is distributed as $\mathcal{N}(\frac{S_i(T_{b(t)})}{1+M_i(T_{b(t)})}, \frac{\sigma^2}{1+M_i(T_{b(t)})})$. Then it is easy to see that

$$\mathbb{P}(\theta_i(t) > \frac{\mu_1 + \mu_i}{2} \mid \mathcal{H}_{t-1}) = Q\left(\frac{\sqrt{1 + M_i(T_{b(t)})}}{\sigma}\left(\frac{\mu_1 + \mu_i}{2} - \frac{S_i(T_{b(t)})}{1 + M_i(T_{b(t)})}\right)\right),$$

which will lead to

$$\mathbb{P}\left(\theta_i(t) > \frac{\mu_1 + \mu_i}{2}, \frac{S_i(T_{b(t)})}{1 + M_i(T_{b(t)})} \leq \frac{\mu_1 + 3\mu_i}{4}, M_i(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right) \leq \frac{1}{T} \tag{52}$$

by an analysis that is almost the same as the one prior to (46). On the other hand, the last summand in (51) satisfies

$$\mathbb{P}\left(\frac{S_i(T_{b(t)})}{1 + M_i(T_{b(t)})} > \frac{\mu_1 + 3\mu_i}{4}, M_i(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right)$$

$$= \mathbb{P}\left(\frac{S_i(T_{b(t)}) - \mu_i M_i(T_{b(t)})}{1 + M_i(T_{b(t)})} > \frac{\mu_1 + 3\mu_i}{4} - \mu_i \frac{M_i(T_{b(t)})}{1 + M_i(T_{b(t)})}, M_i(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right)$$

$$\leq \mathbb{P}\left(\frac{S_i(T_{b(t)}) - \mu_i M_i(T_{b(t)})}{1 + M_i(T_{b(t)})} > \frac{\Delta_i}{8}, M_i(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right).$$

Similar to the analysis in (50), setting $\lambda = \frac{\Delta_i}{2}$ in Lemma 4 leads to

$$\mathbb{P}\left(\frac{S_i(T_{b(t)})}{1 + M_i(T_{b(t)})} > \frac{\mu_1 + 3\mu_i}{4}, M_i(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\right) \leq \frac{1}{T}. \tag{53}$$

In view of (51), combining (52) and (53) finishes the proof of (43). □

D.3 Bounds on Functions of Q

This section provides bounds on various functions of Q function. Before we state our results, we introduce for any i

$$\hat{\tau}_{i,j} = \begin{cases} \min\{t \in \mathbb{Z}^+ | A_t = i, t = C_{b,k} \text{ or } C_{e,k} \text{ for some } k\} & \text{if } j = 1 \\ \min\{t \in \mathbb{Z}^+ | A_t = i, t = C_{b,k} \text{ or } C_{e,k} \text{ for some } k, t > \hat{\tau}_{i,j-1}\} & \text{if } j > 1. \end{cases}$$

Note that if a set is empty, then $\hat{\tau}_{i,j}$ is set to be infinity. Here $\hat{\tau}_{i,j}$ denotes the time index where we choose the ith arm for j^{th} time at the beginning or at the end of a cycle. It is clear that $\hat{\tau}_{i,j}$ is a stopping time for $\{\mathcal{F}_t\}$ specified in Lemma 4, so (31) in Lemma 4 remains true if we set τ to be $\hat{\tau}_{i,j}$:

$$\mathbb{E}[\lambda(S_i(T_{b(\hat{\tau}_{i,j})}) - \mu_i M_i(T_{b(\hat{\tau}_{i,j})})) - \frac{\lambda^2}{8}(1 + M_i(T_{b(\hat{\tau}_{i,j})}))] \leq 1 \tag{54}$$

for any $i \in \{1, 2, ..., K\}$ and real λ.

Lemma 6. *For any $j > 1$ and $i \in \{1, 2, ..., K\}$, if $x \geq 0$:*

$$\mathbb{P}\left(\hat{\tau}_{i,j} < \infty, \frac{S_i(T_{b(\hat{\tau}_{i,j})}) - \mu_i M_i(T_{b(\hat{\tau}_{i,j})})}{\sqrt{1 + M_i(T_{b(\hat{\tau}_{i,j})})}} > x\right) \leq \exp(-2x^2/\alpha), \tag{55}$$

and if $x \leq 0$:

$$\mathbb{P}\left(\hat{\tau}_{i,j} < \infty, \frac{S_i(T_{b(\hat{\tau}_{i,j})}) - \mu_i M_i(T_{b(\hat{\tau}_{i,j})})}{\sqrt{1 + M_i(T_{b(\hat{\tau}_{i,j})})}} < x\right) \leq \exp(-2x^2/\alpha). \tag{56}$$

19

Proof. We start with the proof of (55). Assume $x \geq 0$. By setting $\lambda = \frac{4x}{\sqrt{\alpha j}}$ in (54), we have

$$
\begin{aligned}
1 &\geq \mathbb{E}[\exp(\frac{4x}{\sqrt{\alpha j}}(S_i(T_{b(\hat{\tau}_{i,j})}) - \mu_i M_i(T_{b(\hat{\tau}_{i,j})})) - \frac{2x^2}{\alpha j}(1 + M_i(T_{b(\hat{\tau}_{i,j})})))] \\
&\geq \mathbb{E}[\exp(\frac{4x}{\sqrt{\alpha j}}(S_i(T_{b(\hat{\tau}_{i,j})}) - \mu_i M_i(T_{b(\hat{\tau}_{i,j})})) - \frac{2x^2}{\alpha j}(1 + M_i(T_{b(\hat{\tau}_{i,j})}))) \\
&\qquad\qquad \mathbb{1}\left(\hat{\tau}_{i,j} < \infty, \frac{S_i(T_{b(\hat{\tau}_{i,j})}) - \mu_1 M_i(T_{b(\hat{\tau}_{i,j})})}{\sqrt{1 + M_i(T_{b(\hat{\tau}_{i,j})})}} > x\right)] \\
&\geq \mathbb{E}[\exp(\frac{4x}{\sqrt{\alpha j}}(S_i(T_{b(\hat{\tau}_{i,j})}) - \mu_i M_i(T_{b(\hat{\tau}_{i,j})})) - \frac{2x^2}{\alpha}) \mathbb{1}(\hat{\tau}_{i,j} < \infty, \frac{S_i(T_{b(\hat{\tau}_{i,j})}) - \mu_1 M_i(T_{b(\hat{\tau}_{i,j})})}{\sqrt{1 + M_i(T_{b(\hat{\tau}_{i,j})})}} > x)]
\end{aligned}
$$
$$(57)$$

Note that the last inequality follows from $M_i(T_{b(\hat{\tau}_{i,j})}) \leq j - 1$ since $T_{b(\hat{\tau}_{i,j})}$ is the last batch end point, which is strictly smaller than $\hat{\tau}_{i,j}$. Also by the construction of Algorithm 1, $\max\{1, \lceil \alpha M_i(T_{b(\hat{\tau}_{i,j})}) \rceil\} \geq j$. Given that $j > 1$, $M_i(T_{b(\hat{\tau}_{i,j})}) \geq 1$, which results in $\alpha(1 + M_i(T_{b(\hat{\tau}_{i,j})})) \geq j$ by $\alpha > 1$. Then on $\{\hat{\tau}_{i,j} < \infty, \frac{S_i(T_{b(\hat{\tau}_{i,j})}) - \mu_i M_i(T_{b(\hat{\tau}_{i,j})})}{\sqrt{1 + M_i(T_{b(\hat{\tau}_{i,j})})}} > x\}$, we have

$$
\frac{4x}{\sqrt{\alpha j}}(S_i(T_{b(\hat{\tau}_{i,j})}) - \mu_i M_i(T_{b(\hat{\tau}_{i,j})})) \geq \frac{4x^2}{\alpha}.
$$

Using this inequality inside (57) proves (55). Finally, the proof of (56) will follow the same steps with a single exception: here we set $\lambda = -\frac{4x}{\sqrt{\alpha j}}$ in (54). This finishes the proof. $\qquad \square$

Proposition 7. *Assume $5\sigma^2/4 \geq \alpha$. Then for any positive integer j and $i \geq 2$, we have*

$$
\mathbb{E}\left[\mathbb{1}(\hat{\tau}_{1,j} < \infty)\frac{1}{Q^2\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)\right)}\right] \leq C \tag{58}
$$

and

$$
\mathbb{E}\left[\mathbb{1}(\hat{\tau}_{i,j} < \infty)\frac{1}{Q^2\left(\frac{\sqrt{1+M_i(T_{b(\hat{\tau}_{i,j})})}}{\sigma}\left(\frac{S_i(T_{b(\hat{\tau}_{i,j})})}{1+M_i(T_{b(\hat{\tau}_{i,j})})} - \frac{\mu_1+\mu_i}{2}\right)\right)}\right] \leq C \tag{59}
$$

where C is an absolute constant independent of the system parameters.

Proof. We start with the proof of (58) and the proof of (59) will follow similarly. First note that if $j = 1$, then M_1 and S_1 terms will be zeros, and as a result, the denominator inside the expectation is lower bounded as follows:

$$
Q(\frac{\mu_1 + \mu_i}{2\sigma}) \geq Q(0.5)
$$

since $\mu_i, \mu_1 \leq 1$ and $\sigma^2 \geq 1$. So for $j = 1$, we can choose C to be $1/Q^2(0.5)$.

For $j > 1$ by (22)

$$\mathbb{E}\left[\mathbb{1}(\hat{\tau}_{1,j} < \infty)\frac{1}{Q^2\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)\right)}\right]$$

$$= \int_{x=0}^{\infty}\mathbb{P}\left(1(\hat{\tau}_{1,j} < \infty)\frac{1}{Q^2\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)\right)} > x\right)dx$$

$$\leq 4 + \int_{x=4}^{\infty}\mathbb{P}\left(\hat{\tau}_{1,j} < \infty, Q\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)\right) < \frac{1}{\sqrt{x}}\right)dx$$

$$\leq 4 + \int_{x=4}^{\infty}\mathbb{P}\left(\hat{\tau}_{1,j} < \infty, \frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right) > Q^{-1}(1/\sqrt{x})\right)dx,$$

$$\tag{60}$$

where the last inequality follows from the fact that $Q(\cdot)$ is a decreasing function. Now note that

$$\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})} = \frac{\mu_1+\mu_i}{2} - \mu_1 + \mu_1 - \frac{\mu_1 M_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})} - \frac{S_1(T_{b(\hat{\tau}_{1,j})}) - \mu_1 M_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}$$

$$\leq \frac{1}{1+M_1(T_{b(\hat{\tau}_{1,j})})} - \frac{S_1(T_{b(\hat{\tau}_{1,j})}) - \mu_1 M_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}.$$

The last inequality follows from $\mu_1 \leq 1$. This analysis shows that if $\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}) > \sigma Q^{-1}(1/\sqrt{x})$, then $-\frac{S_1(T_{b(\hat{\tau}_{1,j})}) - \mu_1 M_1(T_{b(\hat{\tau}_{1,j})})}{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}} > \sigma Q^{-1}(1/\sqrt{x}) - \frac{1}{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}$. Since $M_1(T_{b(\hat{\tau}_{1,j})}) \geq 0$ and $\sigma^2 \geq 1$, the last inequality leads to

$$\mathbb{E}\left[\mathbb{1}(\hat{\tau}_{1,j} < \infty)\frac{1}{Q^2\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)\right)}\right]$$

$$\leq 4 + \int_{x=4}^{\infty}\mathbb{P}\left(\hat{\tau}_{1,j} < \infty, -\frac{S_1(T_{b(\hat{\tau}_{1,j})}) - \mu_1 M_1(T_{b(\hat{\tau}_{1,j})})}{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}} > \sigma(Q^{-1}(1/\sqrt{x}) - 1)\right)dx \tag{61}$$

by (60). However, (21) indicates that there exists x_0 such that if $x \geq x_0$, then $Q^{-1}(\frac{1}{\sqrt{x}}) - 1 \geq \sqrt{\frac{16}{25}\log(x)}$, which leads to

$$\mathbb{P}\left(\hat{\tau}_{1,j} < \infty, -\frac{S_1(T_{b(\hat{\tau}_{1,j})}) - \mu_1 M_1(T_{b(\hat{\tau}_{1,j})})}{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}} > \sigma(Q^{-1}(1/\sqrt{x}) - 1)\right) \leq \exp(-\frac{32\sigma^2}{25\alpha}\log(x))$$

by Lemma 6 for $x \geq x_0$. Since $\alpha \leq \frac{5\sigma^2}{4}$, the last inequality can be refined to

$$\mathbb{P}\left(\hat{\tau}_{1,j} < \infty, -\frac{S_1(T_{b(\hat{\tau}_{1,j})}) - \mu_1 M_1(T_{b(\hat{\tau}_{1,j})})}{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}} > \sigma(Q^{-1}(1/\sqrt{x}) - 1)\right) \leq \frac{1}{x^{128/125}} \tag{62}$$

for any $x \geq x_0$. Finally, (61) and (62) shows that

$$\mathbb{E}\left[\mathbb{1}(\hat{\tau}_{1,j} < \infty)\frac{1}{Q^2\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)\right)}\right] \leq x_0 + \int_{x=x_0}^{\infty}\frac{1}{x^{128/125}}dx$$

21

This result proves (58) since x_0 is an absolute constant independent of the system parameters and the final integral is finite. Note that we already upper bounded the $j = 1$ term earlier by an absolute constant so we can just take the maximum of the two.

The proof of (59) will similarly follow. First of all, for $j = 1$ we can bound the Q terms inside the expectation as follows

$$Q(-\frac{\mu_1 + \mu_i}{2\sigma}) \geq Q(0) = 0.5$$

where these follow from the fact that expected means are non-negative. So for $j = 1$, we can choose C to be 4. If $j > 1$, we have

$$\mathbb{E}\left[\mathbb{1}(\hat{\tau}_{i,j} < \infty) \frac{1}{Q^2\left(\frac{\sqrt{1+M_i(T_{b(\hat{\tau}_{i,j})})}}{\sigma} \left(\frac{S_i(T_{b(\hat{\tau}_{i,j})})}{1+M_i(T_{b(\hat{\tau}_{i,j})})} - \frac{\mu_1+\mu_i}{2} \right) \right)} \right]$$

$$\leq 4 + \int_{x=4}^{\infty} \mathbb{P}\left(\hat{\tau}_{i,j} < \infty, \frac{\sqrt{1+M_i(T_{b(\hat{\tau}_{i,j})})}}{\sigma} \left(\frac{S_i(T_{b(\hat{\tau}_{i,j})})}{1+M_i(T_{b(\hat{\tau}_{i,j})})} - \frac{\mu_1+\mu_i}{2} \right) > Q^{-1}(1/\sqrt{x}) \right) dx.$$

$$(63)$$

As for the terms inside:

$$\frac{S_i(T_{b(\hat{\tau}_{i,j})})}{1+M_i(T_{b(\hat{\tau}_{i,j})})} - \frac{\mu_1+\mu_i}{2} = \frac{S_i(T_{b(\hat{\tau}_{i,j})}) - \mu_i M_i(T_{b(\hat{\tau}_{i,j})})}{1+M_i(T_{b(\hat{\tau}_{i,j})})} + \frac{\mu_i M_i(T_{b(\hat{\tau}_{i,j})})}{1+M_i(T_{b(\hat{\tau}_{i,j})})} - \mu_i + \mu_i - \frac{\mu_1+\mu_i}{2}$$

$$\leq \frac{S_i(T_{b(\hat{\tau}_{i,j})}) - \mu_i M_i(T_{b(\hat{\tau}_{i,j})})}{1+M_i(T_{b(\hat{\tau}_{i,j})})},$$

where the last step follows from $\mu_i \geq 0$. This analysis and (63) lead to

$$\mathbb{E}\left[\mathbb{1}(\hat{\tau}_{i,j} < \infty) \frac{1}{Q^2\left(\frac{\sqrt{1+M_i(T_{b(\hat{\tau}_{i,j})})}}{\sigma} \left(\frac{S_i(T_{b(\hat{\tau}_{i,j})})}{1+M_i(T_{b(\hat{\tau}_{i,j})})} - \frac{\mu_1+\mu_i}{2} \right) \right)} \right]$$

$$\leq 4 + \int_{x=4}^{\infty} \mathbb{P}\left(\hat{\tau}_{i,j} < \infty, \frac{S_i(T_{b(\hat{\tau}_{i,j})}) - \mu_i M_i(T_{b(\hat{\tau}_{i,j})})}{\sqrt{1+M_i(T_{b(\hat{\tau}_{i,j})})}} > \sigma Q^{-1}(1/\sqrt{x}) \right) dx.$$

The rest of the proof follows exactly the same way it did after (61) in the proof of (58). □

Proposition 8. *Assume $5\sigma^2/4 \geq \alpha$, $T \geq 2$, and $j > 1024\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2}$, then*

$$\mathbb{E}\left[\mathbb{1}(\hat{\tau}_{1,j} < \infty) \left(\frac{1}{Q\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma} \left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})} \right) \right)} - 1 \right)^2 \right] \leq \frac{C}{T}$$

where C is an absolute constant independent of the system parameters.

Proof. Since $j > 1$, we know that $j \leq \lceil \alpha \times M_1(T_{b(\hat{\tau}_{1,j})}) \rceil$ by the construction of Algorithm 1, and as a result,

$$1024\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2} < j \leq \lceil \alpha \times M_1(T_{b(\hat{\tau}_{1,j})}) \rceil \leq 2\alpha M_1(T_{b(\hat{\tau}_{1,j})}),$$

22

which leads to $M_1(T_{b(\hat{\tau}_{1,j})}) > 512\sigma^2 \frac{\log(T)}{\Delta_i^2}$. Note that $M_1(T_{b(\hat{\tau}_{1,j})}) \geq 1$ if $j > 1$. Then by (22) we have

$$\mathbb{E}\left[\mathbb{1}(\hat{\tau}_{1,j} < \infty)\left(\frac{1}{Q\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)\right)} - 1 \right)^2 \right]$$

$$= \int_{x=0}^{\infty} \mathbb{P}\left(\mathbb{1}(\hat{\tau}_{1,j} < \infty)\left(\frac{1}{Q\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)\right)} - 1 \right)^2 > x \right) dx$$

$$= \int_{x=0}^{\infty} \mathbb{P}\left(\hat{\tau}_{1,j} < \infty, Q\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)\right) < \frac{1}{1+\sqrt{x}} \right) dx$$

$$= \int_{x=0}^{\infty} \mathbb{P}\left(\hat{\tau}_{1,j} < \infty, \frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right) > Q^{-1}(\frac{1}{1+\sqrt{x}}) \right) dx.$$
(64)

These steps follow from simple algebra and the decreasing nature of $Q(\cdot)$.

Now the goal here is to divide the integral in (64) into three regions, where the the contribution from each is in the order of $1/T$. As we will show next, the first region $[0, 1/T]$ will be upper bounded by $1/T$. In the second region $[1/T, x_0]$ for some big x_0, the probability term will be of $O(1/T)$. Finally, the integrand in the third region $[x_0, \infty)$ will decay faster than $\frac{1}{Tx^\delta}$ for some $\delta > 1$, and this will result in a contribution of order $O(1/T)$. The fact that x_0 is an absolute constant will finish the proof. We will start the analysis with the second region. First of all, the symmetry of the normal distribution leads to

$$Q^{-1}(\frac{1}{1+\sqrt{x}}) = -Q^{-1}(\frac{\sqrt{x}}{1+\sqrt{x}})$$
(65)

for $x \geq 0$. Given that $Q(y) \leq \exp(-y^2/2)$ for $y \geq 0$ by the Chernoff bound, we also know $Q^{-1}(\exp(-y^2/2)) \leq y$ for $y \geq 0$. Clearly $\sqrt{2\log(1+\sqrt{T})} \geq 0$ and letting $y = \sqrt{2\log(1+\sqrt{T})}$ shows that

$$Q^{-1}(\frac{\sqrt{x}}{1+\sqrt{x}}) \leq \sqrt{2\log(1+\sqrt{T})}$$

if $x = 1/T$. Since $Q^{-1}(\frac{1}{1+\sqrt{x}})$ is an increasing function of x, for $x \geq 1/T$, we have

$$Q^{-1}(\frac{1}{1+\sqrt{x}}) \geq -\sqrt{2\log(1+\sqrt{T})}.$$

by (65). This analysis proves that

$$Q^{-1}(\frac{1}{1+\sqrt{x}}) + 2\sqrt{2\log(T)} \geq 0$$
(66)

for $x \geq 1/T$ since $T \geq 2$. In addition, by simple algebra

$$\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)$$

$$= \frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \mu_1 + \mu_1 - \frac{\mu_1 M_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})} - \frac{S_1(T_{b(\hat{\tau}_{1,j})}) - \mu_1 M_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)$$

$$= \frac{\mu_1 M_1(T_{b(\hat{\tau}_{1,j})}) - S_1(T_{b(\hat{\tau}_{1,j})})}{\sigma\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}} + \frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(-\frac{\Delta_i}{2} + \frac{\mu_1}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right).$$
(67)

Note here that $\frac{\mu_1}{1+M_1(T_{b(\hat{\tau}_{1,j})})} \leq \frac{\Delta_i^2}{512\sigma^2 \log(T)} \leq \frac{\Delta_i}{4}$ since $M_1(T_{b(\hat{\tau}_{1,j})}) > 512\sigma^2 \frac{\log(T)}{\Delta_i^2}$, $T \geq 2$, and $\sigma^2 \geq 1$. Then (67) leads to

$$\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)$$

$$\leq \frac{\mu_1 M_1(T_{b(\hat{\tau}_{1,j})}) - S_1(T_{b(\hat{\tau}_{1,j})})}{\sigma\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}} - \frac{\Delta_i\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{4\sigma}$$

$$\leq \frac{\mu_1 M_1(T_{b(\hat{\tau}_{1,j})}) - S_1(T_{b(\hat{\tau}_{1,j})})}{\sigma\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}} - 4\sqrt{2\log(T)}. \tag{68}$$

where the last inequality is the result of $M_1(T_{b(\hat{\tau}_{1,j})}) > 512\sigma^2 \frac{\log(T)}{\Delta_i^2}$. The overall analysis shows that if $x \geq 1/T$:

$$\mathbb{P}\left(\hat{\tau}_{1,j} < \infty, \frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right) > Q^{-1}(\frac{1}{1+\sqrt{x}})\right)$$

$$\leq \mathbb{P}\left(\hat{\tau}_{1,j} < \infty, \frac{\mu_1 M_1(T_{b(\hat{\tau}_{1,j})}) - S_1(T_{b(\hat{\tau}_{1,j})})}{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}} > 4\sigma\sqrt{2\log(T)} + \sigma Q^{-1}(\frac{1}{1+\sqrt{x}})\right) \tag{69}$$

$$\leq \exp\left(-\frac{2\sigma^2(4\sqrt{2\log(T)} + Q^{-1}(\frac{1}{1+\sqrt{x}}))^2}{\alpha}\right) \tag{70}$$

$$\leq \exp\left(-\frac{16\sigma^2 \log(T)}{\alpha}\right) \exp\left(-\frac{2\sigma^2(2\sqrt{2\log(T)} + Q^{-1}(\frac{1}{1+\sqrt{x}}))^2}{\alpha}\right) \tag{71}$$

$$\leq \frac{1}{T} \tag{72}$$

(69) follows from (68). Lemma 6 and (66) lead to (70). Similarly (71) is due to (66) and the fact that $(a+b)^2 \geq a^2 + b^2$ for any non-negative a and b. Finally, (72) follows from the fact that $\alpha \leq \frac{5\sigma^2}{4}$. However, for big x values, we can provide a tighter upper bound. Firstly, $Q^{-1}(\frac{1}{1+\sqrt{x}}) \geq \sqrt{\frac{2}{3}\log(x)}$ if $x \geq x_0$ for some $x_0 \geq 4$ by (21). So for $x \geq x_0$

$$\exp\left(-\frac{2\sigma^2(2\sqrt{2\log(T)} + Q^{-1}(\frac{1}{1+\sqrt{x}}))^2}{\alpha}\right) \leq \exp(-\frac{4\sigma^2 \log(x)}{3\alpha}) \leq \frac{1}{x^{16/15}}$$

where we used the fact that $\alpha \leq \frac{5\sigma^2}{4}$. This inequality shows that

$$\mathbb{P}\left(\hat{\tau}_{1,j} < \infty, \frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right) > Q^{-1}(\frac{1}{1+\sqrt{x}})\right) \leq \frac{1}{Tx^{16/15}} \tag{73}$$

if $x \geq x_0$ by (71). Overall, if we plug (72) and (73) back into (64), we have

$$\mathbb{E}\left[\mathbb{1}(\hat{\tau}_{1,j} < \infty)\left(\frac{1}{Q\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)\right)} - 1\right)^2\right] \leq \frac{1+x_0}{T} + \int_{x=x_0}^{\infty} \frac{1}{Tx^{16/15}},$$

which finishes the proof since x_0 is an absolute constant. $\square$

E   Proof of Theorem 1

First of all, (3) is the immediate result of (1) and (2). To prove (4), note that the regret contribution from the arms with $\Delta_i \leq \sigma\sqrt{\frac{\alpha K \log(T)}{T}}$ in the first T rounds can not exceed $\sigma\sqrt{\alpha KT \log(T)}$.

As for any arm with $\Delta_i > \sigma\sqrt{\frac{\alpha K \log(T)}{T}}$, by (2) $\Delta_i \, \mathbb{E}[N_i(T)] \leq \frac{C_1 \sigma \sqrt{\alpha T \log(T)}}{\sqrt{K}}$, which leads to a maximum regret contribution of $C_1 \sigma \sqrt{\alpha K T \log(T)}$ from these arms. As a result, $\mathbb{E}[R(T)] \leq (1 + C_1)\sigma\sqrt{\alpha K T \log(T)}$ and this proves (4).

We will now prove (2). Lets pick any $i \geq 2$. We start the proof by first dividing $N_i(T)$ into smaller terms as follows:

$$\mathbb{E}[N_i(T)] = \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}(A_t = i, \theta_i(t) > \frac{\mu_1 + \mu_i}{2})] + \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}(A_t = i, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2})]$$

$$\leq \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}(A_t = i, \theta_i(t) > \frac{\mu_1 + \mu_i}{2}, M_i(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2})] \tag{74}$$

$$+ \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}(A_t = i, t = C_{e,k} \text{ for some } k, \theta_i(t) > \frac{\mu_1 + \mu_i}{2}, M_i(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})] \tag{75}$$

$$+ \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}(A_t = i, t \neq C_{e,k} \text{ for all } k, \theta_i(t) > \frac{\mu_1 + \mu_i}{2}, M_i(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})] \tag{76}$$

$$+ 1 + \mathbb{E}[\sum_{t=2}^{T} \mathbb{1}(A_t = i, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2}, M_1(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2})] \tag{77}$$

$$+ \mathbb{E}[\sum_{t=2}^{T} \mathbb{1}(A_t = i, A_{t-1} = 1, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2}, M_1(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})] \tag{78}$$

$$+ \mathbb{E}[\sum_{t=2}^{T} \mathbb{1}(A_t = i, A_{t-1} \neq 1, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2}, M_1(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})] \tag{79}$$

Before we proceed to bound each of these terms, we provide our understanding as to how we decomposed this $N_i(T)$ term.

How to Decompose $N_i(T)$ and Handle Each Term: Since Thompson sampling relies on samples $\{\theta_i(t)\}_{i=1}^{K}$ to decide on the next action, it is natural that the term $N_i(T)$ is split whether $\theta_i(t)$ is above a certain threshold or not. In our analysis, we set this threshold to be $\frac{\mu_1 + \mu_i}{2}$ and the resulting terms are $\sum_{t=1}^{T} \mathbb{1}(A_t = i, \theta_i(t) > \frac{\mu_1 + \mu_i}{2})$ and $\sum_{t=1}^{T} \mathbb{1}(A_t = i, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2})$. We should note that as the agent observes more rewards from the ith arm, its samples $\theta_i(t)$ will concentrate around its true mean μ_i and the contribution from $\mathbb{1}(A_t = i, \theta_i(t) > \frac{\mu_1 + \mu_i}{2})$ terms will diminish. On the other hand, if the samples are already near their true mean, then as the agent receives more rewards from the optimal arm, her likelihood of playing the ith arm will decrease, i.e. the contributions from the terms $\mathbb{1}(A_t = i, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2})$ will diminish. Given this discussion, it is natural to center the decomposition of the terms $\sum_{t=1}^{T} \mathbb{1}(A_t = i, \theta_i(t) > \frac{\mu_1 + \mu_i}{2})$, i.e. (74)-(76), and $\sum_{t=1}^{T} \mathbb{1}(A_t = i, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2})$, i.e. (77)-(79), around M_i and M_1 respectively. Earlier discussion also indicates that if there are enough observations from i^{th} and the optimal arm, i.e. M_i and M_1 are above reasonable thresholds, the agent's estimates should be close to the true values, which means that the terms in (74) and (77) should be small, i.e. of order $O(1)$. The main technical difficulty of analyzing the rest of the decomposition terms is that our algorithm uses random stopping times to decide on batch end points. So we need to show that these cycles are relatively short and their lengths do not diverge before the agent receives enough feedback. One way to show that these cycles are short is that we introduce conditions that naturally end the cycles, $\{t = C_{e,k}\}$ in (75) and $\{A_t = i, A_{t-1} = 1\}$ in (78). Since we assume that M_i and M_1 are bounded in these terms, the contributions from (75) and (78) will be of order $O(\log(T))$. As for the terms in (76) and (79), even though it is not immediately clear that cycles containing these elements will be short, we can see that the conditions inside these indicator functions should naturally lead to short cycles. This is because in (76), even if the terms are not cycle end points and there not enough observations from the ith arm, $\{\theta_i(t) > \frac{\mu_1 + \mu_i}{2}\}$ condition can not persist for long. As for the term in (79), the conditions $\{A_{t-1} \neq 1\}$ and $\{\theta_i(t) \leq \frac{\mu_1 + \mu_i}{2}\}$ work against each other from the earlier discussion. As such, most of the proof is dedicated to showing that the terms in (76) and (79) indeed lead to short cycles

on average and they are of order $O(\log(T))$. We achieve this result by using a modified version of Lemma 2.8 [2] and the tails bounds specialized for our algorithm, which are the results from Section D. We now present the technical proof.

First of all, the terms with $M_i(T_{b(t)})$ or $M_1(T_{b(t)})$ being bigger than a certain level account for the estimation error and will be bounded by constants. We know that by (43) of Proposition 5 that

$$\mathbb{E}[\sum_{t=1}^{T} \mathbb{1}(A_t = i, \theta_i(t) > \frac{\mu_1 + \mu_i}{2}, M_i(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2})] \leq 2. \tag{80}$$

Now note that if $A_t = i$ and $\theta_i(t) \leq \frac{\mu_1+\mu_i}{2}$, then $\theta_1(t) \leq \frac{\mu_1+\mu_i}{2}$. That means

$$\mathbb{E}[\sum_{t=2}^{T} \mathbb{1}(A_t = i, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2}, M_1(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2})]$$

$$\leq \mathbb{E}[\sum_{t=2}^{T} \mathbb{1}(\theta_1(t) \leq \frac{\mu_1 + \mu_i}{2}, M_1(T_{b(t)}) \geq 32\sigma^2 \frac{\log(T)}{\Delta_i^2})] \leq 2 \tag{81}$$

where the last inequality follows from (42) of Proposition 5.

Now we bound the remaining terms. We first note that the condition $\{A_t = i, t = C_{e,k} \text{ for some } k\}$ signifies a cycle where the ith arm has been played at the end. Considering that the cycle count from the last batch can not increase more than its α multiple plus one by Algorithm 1, we know

$$\mathbb{E}[\sum_{t=1}^{T} \mathbb{1}(A_t = i, t = C_{e,k} \text{ for some } k, \theta_i(t) > \frac{\mu_1 + \mu_i}{2}, M_i(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})] \leq 32\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2} + 1 \tag{82}$$

due to the $\{M_i(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\}$ condition restricting the number of times we can count a unique cycle. Similarly the condition $\{A_t = i, A_{t-1} = 1\}$ means that $t - 1$ is either a cycle beginning or end point, and again by $M_1(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}$ limiting the number of times we can count a unique cycle with the first action we have

$$\mathbb{E}[\sum_{t=2}^{T} \mathbb{1}(A_t = i, A_{t-1} = 1, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2}, M_1(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})] \leq 32\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2} + 1. \tag{83}$$

Finally the only summands we did not bound are in (76) and (79). We will start with the harder (76) one and use the analysis there to bound (79) at the end. Let $\hat{\mathcal{H}}_t = \{\mathcal{H}_t, \theta_{A_1}(1), \theta_{A_2}(2), ..., \theta_{A_t}(t)\}$ and define the following stopping times for $\{\hat{\mathcal{H}}_t\}$

$$\tau_{b,k} = \begin{cases} \min\{t \in \mathbb{Z}^+ | A_t = i, t \neq C_{e,k} \text{ for all } k, \theta_i(t) > \frac{\mu_1+\mu_i}{2}\} & \text{if } k = 1 \\ \min\{t \in \mathbb{Z}^+ | A_t = i, t \neq C_{e,k} \text{ for all } k, \theta_i(t) > \frac{\mu_1+\mu_i}{2}, t > \tau_{e,k-1}\} & \text{if } k > 1 \end{cases}$$

and

$$\tau_{e,k} = \min\{t \in \mathbb{Z}^+ | A_t \neq i \text{ or } \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2} \text{ such that } t > \tau_{b,k}\}.$$

Note that if any of these min operators are over an empty set, then the random variable is set to infinity. By the definitions of $\tau_{b,k}$ and $\tau_{e,k}$, it is easy to see that $\mathbb{1}(A_t = i, t \neq C_{e,k} \text{ for all } k, \theta_i(t) > \frac{\mu_1+\mu_i}{2}, M_i(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}) = 1$ only if $\tau_{b,k} \leq t < \tau_{e,k}$ for some k. This observations suggests that it is enough to only consider the intervals of $[\tau_{b,k}, \tau_{e,k} - 1]$ while summing over the elements of $1(A_t = i, t \neq C_{e,k} \text{ for all } k, \theta_i(t) > \frac{\mu_1+\mu_i}{2}, M_i(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})$ in (76). However, we can only consider an interval of $[\tau_{b,k}, \tau_{e,k} - 1]$ if $\tau_{b,k} < \infty$. That means we have the following bound

$$\sum_{t=1}^{T} \mathbb{1}(A_t = i, t \neq C_{e,k} \text{ for all } k, \theta_i(t) > \frac{\mu_1 + \mu_i}{2}, M_i(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})$$

$$= \sum_{t=1}^{T} \mathbb{1}(A_t = i, t \neq C_{e,k} \text{ for all } k, \theta_i(t) > \frac{\mu_1 + \mu_i}{2}, M_i(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(t) \leq T)$$

$$\tag{84}$$

$$\leq \sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty)$$

$$\sum_{t=\tau_{b,k}}^{\tau_{e,k}-1} \mathbb{1}(A_t = i, t \neq C_{e,k} \text{ for all } k, \theta_i(t) > \frac{\mu_1 + \mu_i}{2}, M_i(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(t) \leq T)$$

$$\tag{85}$$

$$\leq \sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T)(\tau_{e,k} - \tau_{b,k}), \tag{86}$$

where (84) follows from the fact that $M_1(t) \leq T$ is satisfied for any $t \leq T$. Earlier discussion leads to (85). Finally, note that the time interval $[\tau_{b,k}, \tau_{e,k} - 1]$, where only the ith action is played, is inside a single cycle, so $M_i(T_{b(t)})$ and $M_1(t)$ stay the same for any $t \in [\tau_{b,k}, \tau_{e,k} - 1]$. This observation and ignoring the condition $\{A_t = i, t \neq C_{e,k} \text{ for all } k, \theta_i(t) > \frac{\mu_1+\mu_i}{2}\}$ inside the indicator functions lead to (86). As the for expectation of the summands in (86):

$$\mathbb{E}[\mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T)(\tau_{e,k} - \tau_{b,k})]$$

$$= \mathbb{E}[\mathbb{E}[\mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T)(\tau_{e,k} - \tau_{b,k})|\hat{\mathcal{H}}_{\tau_{b,k}}]]$$

$$= \mathbb{E}[\mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T) \mathbb{E}[\tau_{e,k} - \tau_{b,k}|\hat{\mathcal{H}}_{\tau_{b,k}}]], \tag{87}$$

The last inequality is the result of the indicator function inside the expectation being measurable with respect to $\hat{\mathcal{H}}_{\tau_{b,k}}$. If $\tau_{b,k} < \infty$, then conditioned on $\hat{\mathcal{H}}_{\tau_{b,k}}$ we know that $\tau_{e,k} - \tau_{b,k} - 1$ is a geometric random variable for failures, where the the success probability is $1 - \mathbb{P}(A_{\tau_{b,k}+1} = i, \theta_i(\tau_{b,k} + 1) > \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})$. The reason is that since $\tau_{b,k} < \infty$, we know that $\tau_{e,k} - \tau_{b,k} \geq 1$ and by the definitions of $\tau_{b,k}$ and $\tau_{e,k}$ $[\tau_{b,k}, \tau_{e,k}]$ defines a time interval in a single cycle like we mentioned earlier. These observations show that the sampling process remains the same throughout $[\tau_{b,k}, \tau_{e,k}]$, and conditioned on $\hat{\mathcal{H}}_{\tau_{b,k}}$ $[\tau_{b,k} + 1, \tau_{e,k} - 1]$ is a period of failures if we were to define success as $A_i(t) \neq i$ or $\theta_i(t) \leq \frac{\mu_1+\mu_i}{2}$. The overall analysis proves the following set of inequalities

$$\mathbb{1}(\tau_{b,k} < \infty) \mathbb{E}[\tau_{e,k} - \tau_{b,k}|\hat{\mathcal{H}}_{\tau_{b,k}}] = \mathbb{1}(\tau_{b,k} < \infty) \frac{1}{1 - \mathbb{P}(A_{\tau_{b,k}+1} = i, \theta_i(\tau_{b,k} + 1) > \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}$$

$$\leq \mathbb{1}(\tau_{b,k} < \infty) \frac{1}{\mathbb{P}(\theta_i(\tau_{b,k} + 1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}$$

27

almost surely. Putting this inequality inside (87) and summing the elements like in (86) shows that

$$\mathbb{E}[\sum_{t=1}^{T} \mathbb{1}(A_t = i, t \neq C_{e,k} \text{ for all } k, \theta_i(t) > \frac{\mu_1 + \mu_i}{2}, M_i(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})]$$

$$\leq \mathbb{E}[\sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T) \frac{1}{\mathbb{P}(\theta_i(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})}]$$

$$= \mathbb{E}[\sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T, A_{\tau_{e,k}} = i) \frac{1}{\mathbb{P}(\theta_i(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})}]$$
$$(88)$$

$$+ \mathbb{E}[\sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T, A_{\tau_{e,k}} \neq i) \frac{1}{\mathbb{P}(\theta_i(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})}]$$
$$(89)$$

Note by the earlier analysis we know that on $\{\tau_{b,k} < \infty\}$ $\tau_{e,k}$ is almost surely finite. Then the last equality follows from dividing the terms according to $A_{\tau_{e,k}} = i$ or not. We will first bound the summand in (88). To that end, we now analyze $\mathbb{1}(\tau_{b,k} < \infty) \mathbb{P}(A_{\tau_{e,k}} = 1 | \hat{\mathcal{H}}_{\tau_{b,k}})$. Note that by the earlier analysis we know the sampling distributions remains the same throughout $[\tau_{b,k}, \tau_{e,k}]$, which leads to

$$\mathbb{P}(A_{\tau_{e,k}} = 1 | \hat{\mathcal{H}}_{\tau_{b,k}}) = \frac{\mathbb{P}(A_{\tau_{b,k}+1} = 1 | \hat{\mathcal{H}}_{\tau_{b,k}})}{1 - \mathbb{P}(A_{\tau_{b,k}+1} = i, \theta_i(\tau_{b,k}+1) > \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})} \quad (90)$$

$$\geq \frac{\mathbb{P}(\theta_1(\tau_{b,k}+1) > \frac{\mu_1+\mu_i}{2}, \theta_j(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2} \text{ for all } j \neq 1 | \hat{\mathcal{H}}_{\tau_{b,k}})}{1 - \mathbb{P}(A_{\tau_{b,k}+1} = i, \theta_i(\tau_{b,k}+1) > \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})} \quad (91)$$

$$= \frac{\mathbb{P}(\theta_1(\tau_{b,k}+1) > \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}}) \mathbb{P}(\theta_j(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2} \text{ for all } j \neq 1 | \hat{\mathcal{H}}_{\tau_{b,k}})}{1 - \mathbb{P}(A_{\tau_{b,k}+1} = i, \theta_i(\tau_{b,k}+1) > \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})} \quad (92)$$

on $\{\tau_{b,k} < \infty\}$. Note that since $\tau_{b,k} < \infty$, all the conditional probabilities stated here are almost surely positive. Here (90) trivially follows from the definition of success and failure of the geometric random variable we have defined earlier, i.e. $\tau_{e,k} - \tau_{b,k} - 1$. (91) is the result of the action selection process where we know that for the first action to be chosen the sample θ_1 has to be at least as big as the other samples. Finally, conditioned on $\hat{\mathcal{H}}_{\tau_{b,k}}$, $\{\theta_j(\tau_{b,k}+1)\}_{j=1}^{K}$ are independent, which results in (92). On the other hand

$$\mathbb{P}(A_{\tau_{e,k}} = i | \hat{\mathcal{H}}_{\tau_{b,k}}) = \frac{\mathbb{P}(A_{\tau_{b,k}+1} = i, \theta_i(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})}{1 - \mathbb{P}(A_{\tau_{b,k}+1} = i, \theta_i(\tau_{b,k}+1) > \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})} \quad (93)$$

$$\leq \frac{\mathbb{P}(\theta_j(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2} \text{ for all } j | \hat{\mathcal{H}}_{\tau_{b,k}})}{1 - \mathbb{P}(A_{\tau_{b,k}+1} = i, \theta_i(\tau_{b,k}+1) > \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})} \quad (94)$$

$$= \frac{\mathbb{P}(\theta_1(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}}) \mathbb{P}(\theta_j(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2} \text{ for all } j \neq 1 | \hat{\mathcal{H}}_{\tau_{b,k}})}{1 - \mathbb{P}(A_{\tau_{b,k}+1} = i, \theta_i(\tau_{b,k}+1) > \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})} \quad (95)$$

on $\{\tau_{b,k} < \infty\}$. Note that (93) follows from the fact that $A_{\tau_{e,k}} = i$ only if $\theta_i(\tau_{e,k}) \leq \frac{\mu_1+\mu_i}{2}$ by the definition of $\tau_{e,k}$. Considering that $\{A_{\tau_{b,k}+1} = i, \theta_i(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2}\}$ means $\theta_j(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2}$ for all j, we have (94). (95) is the result of the conditional independence. Combining (92) and (95) leads to

$$\mathbb{P}(A_{\tau_{e,k}} = i | \hat{\mathcal{H}}_{\tau_{b,k}}) \leq \frac{\mathbb{P}(\theta_1(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})}{\mathbb{P}(\theta_1(\tau_{b,k}+1) > \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})} \mathbb{P}(A_{\tau_{e,k}} = 1 | \hat{\mathcal{H}}_{\tau_{b,k}}). \quad (96)$$

on $\{\tau_{b,k} < \infty\}$. As a result

$$\mathbb{E}[\mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T, A_{\tau_{e,k}} = i)\frac{1}{\mathbb{P}(\theta_i(\tau_{b,k} + 1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}]$$

$$= \mathbb{E}[\mathbb{E}[\mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T)\frac{1}{\mathbb{P}(\theta_i(\tau_{b,k} + 1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}$$

$$\mathbb{1}(A_{\tau_{e,k}} = i)|\hat{\mathcal{H}}_{\tau_{b,k}}]]$$

$$= \mathbb{E}[\mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T)\frac{1}{\mathbb{P}(\theta_i(\tau_{b,k} + 1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}$$

$$\mathbb{P}(A_{\tau_{e,k}} = i|\hat{\mathcal{H}}_{\tau_{b,k}})] \qquad (97)$$

$$\leq \mathbb{E}[\mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T)\frac{1}{\mathbb{P}(\theta_i(\tau_{b,k} + 1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}$$

$$\frac{\mathbb{P}(\theta_1(\tau_{b,k} + 1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}{\mathbb{P}(\theta_1(\tau_{b,k} + 1) > \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}\mathbb{P}(A_{\tau_{e,k}} = 1|\hat{\mathcal{H}}_{\tau_{b,k}})] \qquad (98)$$

$$= \mathbb{E}[\mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T)\frac{1}{\mathbb{P}(\theta_i(\tau_{b,k} + 1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}$$

$$\frac{\mathbb{P}(\theta_1(\tau_{b,k} + 1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}{\mathbb{P}(\theta_1(\tau_{b,k} + 1) > \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}\mathbb{1}(A_{\tau_{e,k}} = 1)]$$

$$\leq \mathbb{E}[\mathbb{1}(\tau_{b,k} < \infty, A_{\tau_{e,k}} = 1, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T)\frac{1}{\mathbb{P}^2(\theta_i(\tau_{b,k} + 1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}]$$

$$+ \mathbb{E}[\mathbb{1}(\tau_{b,k} < \infty, A_{\tau_{e,k}} = 1, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T)\frac{\mathbb{P}^2(\theta_1(\tau_{b,k} + 1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}{\mathbb{P}^2(\theta_1(\tau_{b,k} + 1) > \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}].$$

$$(99)$$

(97) follows from the measurability of the indicator function and the inverse probability term with respect to $\hat{\mathcal{H}}_{\tau_{b,k}}$. (96) leads to (98). Finally, the last inequality follows from the fact that $2\sqrt{a \times b} \leq a + b$ for any non-negative a and b. Here (99) shows that

$$\mathbb{E}[\sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T, A_{\tau_{e,k}} = i)\frac{1}{\mathbb{P}(\theta_i(\tau_{b,k} + 1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}]$$

$$\leq \mathbb{E}[\sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty, A_{\tau_{e,k}} = 1, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})\frac{1}{\mathbb{P}^2(\theta_i(\tau_{b,k} + 1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}]$$

$$+ \mathbb{E}[\sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty, A_{\tau_{e,k}} = 1, M_1(\tau_{b,k}) \leq T)\frac{\mathbb{P}^2(\theta_1(\tau_{b,k} + 1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}{\mathbb{P}^2(\theta_1(\tau_{b,k} + 1) > \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}] \qquad (100)$$

Here we eliminated one condition from each indicator function in the last inequality. However, we know by the action selection process of Thompson sampling and $b(\tau_{b,k} + 1) = b(\tau_{b,k})$ equality due to $\tau_{b,k}$ and $\tau_{b,k} + 1$ being in the same cycle that

$$\mathbb{P}(\theta_1(\tau_{b,k} + 1) > \frac{\mu_1 + \mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}}) = Q\left(\frac{\sqrt{1 + M_1(T_{b(\tau_{b,k})})}}{\sigma}\left(\frac{\mu_1 + \mu_i}{2} - \frac{S_1(T_{b(\tau_{b,k})})}{1 + M_1(T_{b(\tau_{b,k})})}\right)\right)$$

$$(101)$$

and

$$\mathbb{P}(\theta_i(\tau_{b,k} + 1) \leq \frac{\mu_1 + \mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}}) = Q\left(\frac{\sqrt{1 + M_i(T_{b(\tau_{b,k})})}}{\sigma}\left(\frac{S_i(T_{b(\tau_{b,k})})}{1 + M_i(T_{b(\tau_{b,k})})} - \frac{\mu_1 + \mu_i}{2}\right)\right)$$

$$(102)$$

29

on $\{\tau_{b,k} < \infty\}$. Considering (101) and (102), we see can view (100) as

$$\mathbb{E}[\sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T, A_{\tau_{e,k}} = i) \frac{1}{\mathbb{P}(\theta_i(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})}]$$

$$\leq \mathbb{E}[\sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty, A_{\tau_{e,k}} = 1, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}) f_1(M_i(T_{b(\tau_{b,k})}), S_i(T_{b(\tau_{b,k})}))]$$

$$+ \mathbb{E}[\sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty, A_{\tau_{e,k}} = 1, M_1(\tau_{b,k}) \leq T) f_2(M_1(T_{b(\tau_{b,k})}), S_1(T_{b(\tau_{b,k})}))]$$

where f_1 and f_2 are some functions with domain $\mathbb{R}_{\geq 0} \times \mathbb{R}$. If we let

$$\hat{\tau}_{i,j} = \begin{cases} \min\{t \in \mathbb{Z}^+ | A_t = i, t = C_{b,k} \text{ or } C_{e,k} \text{ for some } k\} & \text{if } j = 1 \\ \min\{t \in \mathbb{Z}^+ | A_t = i, t = C_{b,k} \text{ or } C_{e,k} \text{ for some } k, t > \hat{\tau}_{i,j-1}\} & \text{if } j > 1, \end{cases}$$

which denotes the time indices where we choose the ith arm at the beginning or at the end of a cycle, we notice that $M_i(T_{b(\tau_{b,k})}) = M_i(T_{b(\hat{\tau}_{i,j})})$ and $S_i(T_{b(\tau_{b,k})}) = S_i(T_{b(\hat{\tau}_{i,j})})$ for some j on $\{\tau_{b,k} < \infty\}$ since $\tau_{b,k} < \infty$ means that the agent has played the ith action at the beginning of the cycle containing $\tau_{b,k}$. However, when we look $\mathbb{1}(\tau_{b,k} < \infty, A_{\tau_{e,k}} = 1, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}) f_1(M_i(T_{b(\tau_{b,k})}), S_i(T_{b(\tau_{b,k})}))$ terms, we realize that each time interval $[\tau_{b,k}, \tau_{e,k} - 1]$ will belong to a different cycle due to $A_{\tau_{e,k}} = 1$, and the condition $\{M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\}$ implies that the indicator function can be non-zero only if $[\tau_{b,k}, \tau_{e,k}]$ is inside the one of the first $\lceil 32\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2} \rceil$ cycles of the ith arm. The overall discussion leads to the following bound:

$$\sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty, A_{\tau_{e,k}} = 1, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}) f_1(M_i(T_{b(\tau_{b,k})}), S_i(T_{b(\tau_{b,k})}))$$

$$\leq \sum_{j=1}^{\lceil 32\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2} \rceil} \mathbb{1}(\hat{\tau}_{i,j} < \infty) f_1(M_i(T_{b(\hat{\tau}_{i,j})}), S_i(T_{b(\hat{\tau}_{i,j})})). \tag{103}$$

On the other hand, on $\{\tau_{b,k} < \infty, A_{\tau_{e,k}} = 1\}$, $M_1(T_{b(\tau_{b,k})}) = M_1(T_{b(\hat{\tau}_{1,j})})$ and $S_1(T_{b(\tau_{b,k})}) = S_1(T_{b(\hat{\tau}_{1,j})})$ for some j since $\tau_{e,k}$ here is the cycle end point. Similar to the earlier analysis, for each k that satisfies the $\{\tau_{b,k} < \infty, A_{\tau_{e,k}} = 1, M_1(\tau_{b,k}) \leq T\}$ condition, $[\tau_{b,k}, \tau_{e,k}]$ will be in a distinct cycle from the first $T + 1$ ones containing the first arm. This observation shows that

$$\sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty, A_{\tau_{e,k}} = 1, M_1(\tau_{b,k}) \leq T) f_2(M_1(T_{b(\tau_{b,k})}), S_1(T_{b(\tau_{b,k})}))$$

$$\leq \sum_{j=1}^{T+1} \mathbb{1}(\hat{\tau}_{1,j} < \infty) f_2(M_1(T_{b(\hat{\tau}_{1,j})}), S_1(T_{b(\hat{\tau}_{1,j})})) \tag{104}$$

In view of (100), (103) and (104) result in the following bound

$$\mathbb{E}[\sum_{k=1}^{\infty} \mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T, A_{\tau_{e,k}} = i) \frac{1}{\mathbb{P}(\theta_i(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2} | \hat{\mathcal{H}}_{\tau_{b,k}})}]$$

$$\leq \mathbb{E}\left[\sum_{j=1}^{\lceil 32\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2} \rceil} \mathbb{1}(\hat{\tau}_{i,j} < \infty) \frac{1}{Q^2\left(\frac{\sqrt{1+M_i(T_{b(\hat{\tau}_{i,j})})}}{\sigma} \left(\frac{S_i(T_{b(\hat{\tau}_{i,j})})}{1+M_i(T_{b(\hat{\tau}_{i,j})})} - \frac{\mu_1+\mu_i}{2}\right)\right)} \right]$$

$$+ \mathbb{E}\left[\sum_{j=1}^{T+1} \mathbb{1}(\hat{\tau}_{1,j} < \infty) \left(\frac{1}{Q\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma} \left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)\right)} - 1\right)^2 \right]. \tag{105}$$

30

where we replaced f_1 and f_2 with their exact forms. Note that although we did not define f_1 and f_2 explicitly, it is easy understand their exact formulation from the earlier discussion, i.e. from the conditional probability functions in (100) and the equalities stated in (101) and (102). Here we know that the first expectation to the right-side of the inequality is upper bounded by $C(1 + 32\alpha\sigma^2\frac{\log(T)}{\Delta_i^2})$ by Proposition 7, where C is an absolute constant. On the other hand, we have

$$\mathbb{E}\left[\mathbb{1}(\hat{\tau}_{1,j} < \infty)\left(\frac{1}{Q\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)\right)} - 1\right)^2\right]$$

$$\leq \begin{cases} C & \text{if } j \leq 1024\alpha\sigma^2\frac{\log(T)}{\Delta_i^2} \\ C/T & \text{if } j > 1024\alpha\sigma^2\frac{\log(T)}{\Delta_i^2}, \end{cases}$$

where C is an absolute constant independent of the system variables. Note that the constant bound follows from Proposition 7, while $O(1/T)$ bound is the result of Proposition 8. This overall analysis and (105) lead to

$$\mathbb{E}[\sum_{k=1}^{\infty}\mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2\frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T, A_{\tau_{e,k}} = i)\frac{1}{\mathbb{P}(\theta_i(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}]$$

$$\leq C(2 + 1056\alpha\sigma^2\frac{\log(T)}{\Delta_i^2}) \tag{106}$$

for an absolute constant C. This proof bounds the term in (88). However, with the analysis we have done so far, bounding the term in (89) is almost immediate. First note that, similar to the earlier analysis, $M_i(T_{b(\tau_{b,k})}) = M_i(T_{b(\hat{\tau}_{i,j})})$ and $S_i(T_{b(\tau_{b,k})}) = S_i(T_{b(\hat{\tau}_{i,j})})$ for some j on $\{\tau_{b,k} < \infty\}$ since $\tau_{b,k} < \infty$ means that the agent has played the ith action at the beginning of the cycle containing $\tau_{b,k}$. Then by (102)

$$\mathbb{P}(\theta_i(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}}) = Q\left(\frac{\sqrt{1+M_i(T_{b(\hat{\tau}_{i,j})})}}{\sigma}\left(\frac{S_i(T_{b(\hat{\tau}_{i,j})})}{1+M_i(T_{b(\hat{\tau}_{i,j})})} - \frac{\mu_1+\mu_i}{2}\right)\right)$$

on $\{\tau_{b,k} < \infty\}$ for some j. However, for each k that satisfies the condition $\{\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2\frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T, A_{\tau_{e,k}} \neq i\}$, $[\tau_{b,k}, \tau_{e,k}]$ has to be in a distinct cycle from the first $\lceil 32\alpha\sigma^2\frac{\log(T)}{\Delta_i^2}\rceil$ cycles of the ith arm. Note that the distinctiveness follows from the fact that $A_{\tau_{e,k}} \neq i$ condition ends the cycle, while the upper bound on the number of cycles is the result of $\{M_i(T_{b(\tau_{b,k})}) < 32\sigma^2\frac{\log(T)}{\Delta_i^2}\}$ and the way Algorithm 1 is implemented. These arguments naturally lead to

$$\mathbb{E}[\sum_{k=1}^{\infty}\mathbb{1}(\tau_{b,k} < \infty, M_i(T_{b(\tau_{b,k})}) < 32\sigma^2\frac{\log(T)}{\Delta_i^2}, M_1(\tau_{b,k}) \leq T, A_{\tau_{e,k}} \neq i)\frac{1}{\mathbb{P}(\theta_i(\tau_{b,k}+1) \leq \frac{\mu_1+\mu_i}{2}|\hat{\mathcal{H}}_{\tau_{b,k}})}]$$

$$\leq \mathbb{E}\left[\sum_{j=1}^{\lceil 32\alpha\sigma^2\frac{\log(T)}{\Delta_i^2}\rceil}\mathbb{1}(\hat{\tau}_{i,j} < \infty)\frac{1}{Q\left(\frac{\sqrt{1+M_i(T_{b(\hat{\tau}_{i,j})})}}{\sigma}\left(\frac{S_i(T_{b(\hat{\tau}_{i,j})})}{1+M_i(T_{b(\hat{\tau}_{i,j})})} - \frac{\mu_1+\mu_i}{2}\right)\right)}\right]$$

$$\leq C(1 + 32\alpha\sigma^2\frac{\log(T)}{\Delta_i^2}) \tag{107}$$

where the last inequality follows from Proposition 7 and the range of Q being from zero to one. In view of (88) and (89), combining (106) and (107) shows that

$$\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}(A_t = i, t \neq C_{e,k} \text{ for all } k, \theta_i(t) > \frac{\mu_1+\mu_i}{2}, M_i(T_{b(t)}) < 32\sigma^2\frac{\log(T)}{\Delta_i^2})]$$

$$\leq C(3 + 1088\alpha\sigma^2\frac{\log(T)}{\Delta_i^2}) \tag{108}$$

31

where C is an absolute constant. This finishes the analysis of the summand in (76).

Finally, we will bound the summand in (79). However, most of the proof ideas will follow from earlier analysis. First note that if $\theta_1(t) > \frac{\mu_1 + \mu_i}{2}$, while $\theta_j(t) \leq \frac{\mu_1 + \mu_i}{2}$ for $j \geq 2$, then $A_t = 1$:

$$\mathbb{P}(A_t = 1 \mid \mathcal{H}_{t-1}) \geq \mathbb{P}(\theta_1(t) > \frac{\mu_1 + \mu_i}{2} \mid \mathcal{H}_{t-1}) \, \mathbb{P}(\theta_j(t) \leq \frac{\mu_1 + \mu_i}{2} \text{ for all } j \neq 1 \mid \mathcal{H}_{t-1}) \quad (109)$$

where we also used the conditional independence of $\theta_j(t)$s given $\mathcal{H}_{t-1}$. On the other hand, if $A_t = i$ and $\theta_i(t) \leq \frac{\mu_1 + \mu_i}{2}$, then $\theta_j(t) \leq \frac{\mu_1 + \mu_i}{2}$ for all $j \geq 2$:

$$\mathbb{P}(A_t = i, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2} \mid \mathcal{H}_{t-1}) \leq \mathbb{P}(\theta_j(t) \leq \frac{\mu_1 + \mu_i}{2} \text{ for all } j \neq 1 \mid \mathcal{H}_{t-1}). \quad (110)$$

The combination of (109) and (110) lead to

$$\mathbb{P}(A_t = i, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2} \mid \mathcal{H}_{t-1}) \leq \frac{\mathbb{P}(A_t = 1 \mid \mathcal{H}_{t-1})}{\mathbb{P}(\theta_1(t) > \frac{\mu_1+\mu_i}{2} \mid \mathcal{H}_{t-1})}. \quad (111)$$

Note that considering the action selection process of Algorithm 1 where conditioned on the past observations θ_1 has a Gaussian distribution, $\mathbb{P}(\theta_1(t) > \frac{\mu_1+\mu_i}{2} \mid \mathcal{H}_{t-1})$ will almost surely be non-zero. Then we have

$$\mathbb{E}[\mathbb{1}(A_t = i, A_{t-1} \neq 1, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2}, M_1(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})]$$

$$= \mathbb{E}[\mathbb{E}[\mathbb{1}(A_t = i, A_{t-1} \neq 1, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2}, M_1(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}) \mid \mathcal{H}_{t-1}]]$$

$$= \mathbb{E}[\mathbb{P}(A_t = i, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2} \mid \mathcal{H}_{t-1}) \, \mathbb{1}(A_{t-1} \neq 1, M_1(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})] \quad (112)$$

$$\leq \mathbb{E}[\frac{\mathbb{P}(A_t = 1 \mid \mathcal{H}_{t-1})}{\mathbb{P}(\theta_1(t) > \frac{\mu_1+\mu_i}{2} \mid \mathcal{H}_{t-1})} \, \mathbb{1}(A_{t-1} \neq 1, M_1(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})] \quad (113)$$

$$= \mathbb{E}[\frac{1}{\mathbb{P}(\theta_1(t) > \frac{\mu_1+\mu_i}{2} \mid \mathcal{H}_{t-1})} \, \mathbb{1}(A_t = 1, A_{t-1} \neq 1, M_1(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})], \quad (114)$$

where (112) follows from moving terms measurable with respect to $\mathcal{H}_{t-1}$ out of the conditional expectation. The bound in (111) leads to (113). Finally, using (114) in (79) shows that

$$\mathbb{E}[\sum_{t=2}^{T} \mathbb{1}(A_t = i, A_{t-1} \neq 1, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2}, M_1(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})]$$

$$\leq \mathbb{E}[\sum_{t=2}^{T} \frac{1}{\mathbb{P}(\theta_1(t) > \frac{\mu_1+\mu_i}{2} \mid \mathcal{H}_{t-1})} \, \mathbb{1}(A_t = 1, A_{t-1} \neq 1, M_1(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})] \quad (115)$$

Note here that if $A_t = 1$ and $A_{t-1} \neq 1$, then t is either a cycle beginning or a cycle end point, which means that $t = \hat{\tau}_{1,j}$ for some j:

$$\mathbb{P}(\theta_1(t) > \frac{\mu_1 + \mu_i}{2} \mid \mathcal{H}_{t-1}) = Q\left(\frac{\sqrt{1 + M_1(T_{b(t)})}}{\sigma} \left(\frac{\mu_1 + \mu_i}{2} - \frac{S_1(T_{b(t)})}{1 + M_1(T_{b(t)})} \right) \right) \quad (116)$$

$$= Q\left(\frac{\sqrt{1 + M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma} \left(\frac{\mu_1 + \mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1 + M_1(T_{b(\hat{\tau}_{1,j})})} \right) \right) \quad (117)$$

where the fact that conditioned on $\mathcal{H}_{t-1}$, $\theta_1(t) \sim \mathcal{N}(\frac{S_1(T_{b(t)})}{1+M_1(T_{b(t)})}, \frac{\sigma^2}{1+M_1(T_{b(t)})})$ leads to (116). In addition, each t that satisfies $\{A_t = 1, A_{t-1} \neq 1, M_1(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2}\}$ condition has to belong to a different cycle and the index j can not be bigger than $\lceil 32\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2} \rceil$. So, in view of (115), (117)
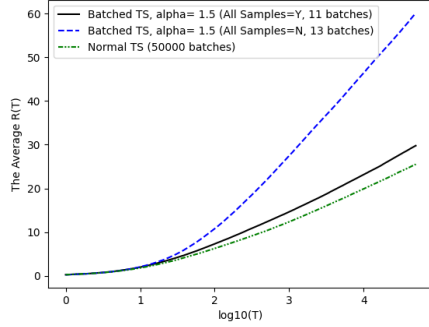
32

leads to

$$\mathbb{E}[\sum_{t=2}^{T} \mathbb{1}(A_t = i, A_{t-1} \neq 1, \theta_i(t) \leq \frac{\mu_1 + \mu_i}{2}, M_1(T_{b(t)}) < 32\sigma^2 \frac{\log(T)}{\Delta_i^2})]$$

$$\leq \mathbb{E}\left[\sum_{j=1}^{\lceil 32\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2}\rceil} \mathbb{1}(\hat{\tau}_{1,j} < \infty) \frac{1}{Q\left(\frac{\sqrt{1+M_1(T_{b(\hat{\tau}_{1,j})})}}{\sigma}\left(\frac{\mu_1+\mu_i}{2} - \frac{S_1(T_{b(\hat{\tau}_{1,j})})}{1+M_1(T_{b(\hat{\tau}_{1,j})})}\right)\right)} \right]$$

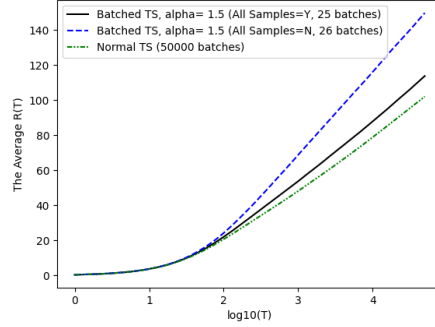$$\leq C(1 + 32\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2}). \tag{118}$$

Here the last inequality is the application of Proposition 7. However, this result finishes the proof of (2) since the collection of bounds, (80), (81), (82), (83), (108), (118), prove that

$$\mathbb{E}[N_i(T)] \leq 6 + 64\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2} + C(4 + 1120\alpha\sigma^2 \frac{\log(T)}{\Delta_i^2}).$$
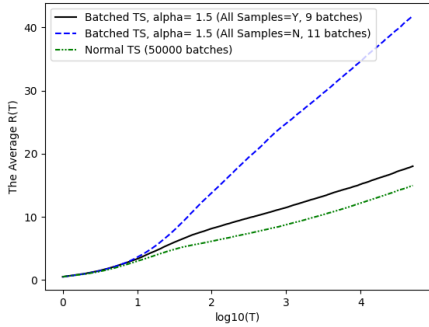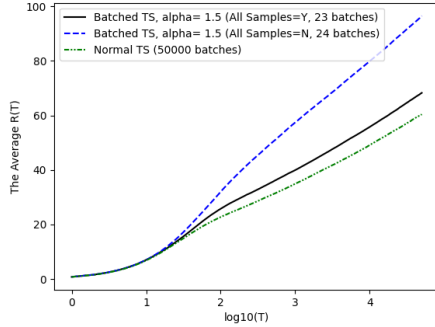
F   Additional Experiments

(a) $Y_{1,t} \sim \text{Bern}(0.75), Y_{2,t} \sim \text{Bern}(0.25)$

(b) $Y_{1,t} \sim \text{Bern}(0.75), Y_{i,t} \sim \text{Bern}(0.25) \; 2 \leq i \leq 5$

(c) $Y_{1,t} \sim \mathcal{N}(1,1), Y_{2,t} \sim \mathcal{N}(0,1)$

(d) $Y_{1,t} \sim \mathcal{N}(1,1), Y_{i,t} \sim \mathcal{N}(0,1) \; 2 \leq i \leq 5$

Figure 2: Empirical Regret Performance of Batched and normal Thompson sampling

In addition to the figures in Section 5, here we provide experiments that showcase how Algorithm 1 that does not make use of all reward observations (shown as All Samples=N) performs compared to the version that uses all the samples (shown as All Samples=Y). Additionally, instead of the average batch count, we report the 95% percentile batch complexity required for each algorithm. As can be seen from these figures, the version of Batched Thompson sampling using all the samples

significantly outperforms the version that does not make use of all the observations. On the other hand, we note that the algorithm that does skip samples (All Samples=N) may require exponentially fewer observations on average than the other version (All Samples=Y). For example, in Figure 2 (a), Algorithm 1 (All Samples=N) uses around 400 samples on average for each of the two arms, while the batched TS algorithm that uses all the samples (All Samples=Y) uses 5×10^4 samples in total. Overall, this indicates a trade-off between fewer samples and possibly lower computation time versus better empirical regret performance.