Asymptotically Efficient Distributed Experimentation

Ilan Lobel, Ankur Mani, Josh Reed

Abstract

Sequential decision making by a large set of myopic agents has gained significant attention over the past decade. In such settings, even a little amount of experimentation from a few agents would benefit all others but obtaining such experimentation could be challenging for a central planner. The academic literature has focused on mechanisms for promoting experimentation through monetary incentives and persuasion through careful information disclosure. In this paper, we study a simple control that the central planner can use to coordinate experimentation. We consider a set of myopic agents that observe their own histories but not the histories of other agents. In a continuous-time stochastic multi-armed bandit model, the agents pick arms myopically and receive instantaneous rewards. Meanwhile, the central planner can observe the history of all agents. We consider a class of policies where the central planner is allowed to irrevocably remove arms. We show that an appropriately chosen policy within this class can generate the needed experimentation and match the regret bounds for a centralized problem thus mitigating the cost of decentralization. We also quantify the minimum number of agents that are needed for such a policy to be asymptotically optimal and the impact of the number of agents on the speed of learning.

1. Introduction

Centralized experimentation and learning has been widely studied in many fields including operations research, economics, computer science and statistics, with multi-armed bandits being the commonly used framework. Over the past decade, there has also been a growing interest in distributed settings, where a central planner desires experimentation but the agents who make decisions are myopic and not interested in experimenting. We study a setting where agents are long-lived and know only the payoff from the actions they took, not the payoffs or the actions chosen by others. The central planner, meanwhile, has access to the entire history of actions and payoffs but has only very minimal control over the system. We provide two examples below.

An important example scenario that has large social value has emerged due to the advent of the Right To Try Act, signed into law in the United States in 2018 [Trickett et al., 2017]. The Right To Try Act enables patients with life-threatening conditions to bypass clinical trials and try new investigational drugs in consultation with their physicians directly. The investigational drugs must have passed the Phase I trials to verify the safety of the drugs on a small population but have not yet gone through Phase II trials to evaluate their effectiveness. Physicians are oath-bound to prescribe the drug that they expect to provide the best outcome for each patient. That is, the physicians are required to be myopic and are unable to experiment with the patients even though

the outcome of an experiment could provide valuable information for future patients. The drug companies and the public health agencies such as the Food and Drug Administration (FDA), are provided the information about the outcomes of the patients for each investigational drug but the physicians often only observe their own actions and outcomes. The practice of trying investigational drugs through the Right To Try Act provides an opportunity to learn their effectiveness outside the standard framework of clinical trials, which are typically lengthy and costly, and has the potential to bring drugs to market faster. But a key challenge for the drug companies and the FDA in doing so is how to generate sufficient experimentation when the physicians who make treatment decisions are myopic.

In such a scenario, the sales agents prefer selling tried-and-tested products which have performed well in their own experiences, rather than experiment with newer, potentially better ones. The upstream firm usually gets feedback about efforts and sales of different products from all agents and is much more informed. The upstream firm would like experimentation to take place but has very little control over which products their downstream agents are trying to sell.

In scenarios like the ones described above, agents are likely to do little to no experimentation due to a lack of incentive, medical/ethical constraints, a lack of patience, or risk aversion. Yet, experimentation in these distributed decision-making environments is critical for learning and long-term success. This dilemma represents the problem of distributed experimentation. This problem has received plenty of attention in the literature recently with proposed solutions involving monetary incentives for experimentation [Frazier et al., 2014], the design of experimentation contracts [Halac et al., 2016], and information design to persuade agents to experiment [Kremer et al., 2014, Che and Hörner, 2018, Mansour et al., 2020].

All of these methods from the literature work in some contexts but not others. For instance, in the right-to-try example, attempting to increase experimentation via payments or contract design is unlikely to be an acceptable option. In this paper, we propose a minimalist control where the central planner is only allowed to remove irrevocably available options for the agents. Removing underperforming products or drugs from the lineup is a method almost always available to the central planner, and we aim to understand its power to drive distributed experimentation. For instance, under the Right To Try Act, eliminating a drug can be implemented by broadcasting that the drug is deemed not sufficiently effective or stopping production. Our proposed policy would achieve fast learning, thus potentially speeding up the drug experimentation and approval phase, while also improving the aggregate outcome for the patients trying the drugs.

To study this distributed experimentation problem and to quantify the power of the proposed method, we present a continuous-time multi-armed bandit model where each bandit arm in the model has an unknown and unique expected reward rate. There are a finite number of agents and at each instant each agent chooses an arm to pull and earns a reward. The cumulative reward collected by all agents who pull a given arm evolve according to independent Brownian motions with

the same unknown drift. The agents observe their past history and are myopic in their decision-making, so they always select the arm which they believe has the highest drift. Thus, the agents are only interested in the exploitation of information for immediate reward and have no interest in experimentation. The central planner observes all of the agent histories and is interested in the total long-term cumulative reward of all agents over a finite time-horizon.

If the central planner had full control over the agents, then the model would reduce to the single-agent continuous-time multi-armed bandit model. For finite time horizon, in order to maximize total reward, the central planner must balance experimentation with exploitation. The central planner's regret is the difference between its reward in the presence of full information about the expected reward rates of the arms and its earned reward in the absence of this information. In the discrete-time version, the expected amount of time an optimal policy of the central planner spends on suboptimal arms is logarithmic in the length of the time-horizon. For completeness, we show that the same rate applies to the continuous-time model. In a multi-armed bandit model with J arms, the minimum expected regret over a time-horizon of T is $\Theta(J \log T)$ (see Lai and Robbins [1985]). The expected amount of experimentation to achieve the minimum regret is also $\Theta(J \log T)$. This can be achieved in the centralized setting through a strategy that plays each arm at least $\Omega(\log T)$ (Lai and Robbins [1985], Auer et al. [2002]).

In the distributed setting, when the central planner has no control over the myopic agents, the above regret serves as a lower bound on the central planner's regret. The class of policies we consider only makes irrevocable decisions of removing arms from all agents. The first question we tackle is what is the cost of myopia and decentralization in this setting. We identify the exact regret for myopic agents without a central planner and show that when there is only one agent then the central planner cannot improve the regret. Through this exercise we also develop new tools to study the continuous time bandit problem and use them to study the setting with multiple agents. We then show that as the time horizon increases, the central planner needs increasingly more agents. If the number of agents grows at a rate $o(\log T)$, that is, at a rate slower than $\log T$, then the expected regret of the central planner grows linearly in the length of the time horizon irrespective of the central planner's policy. We conclude from this result that the long-term cost of decentralization can be high.

This leads us to the second question of whether there is a policy that produces sufficient experimentation if the number of agents is $\Theta(\log T)$. Dropping arms might not seem like a way to generate experimentation, but it can serve this purpose by getting agents which are pulling suboptimal arms to try new ones. We study whether the policy of removing arms once they are deemed to be suboptimal with high confidence can produce sufficient experimentation. The policy we propose involves the central planner, using the historical information from all agents, maintaining a non-decreasing anchor rate at all times which is the minimum expected reward rate it wants all agents to generate in the future. It also maintains confidence bounds for the expected rewards rates

of different arms. These confidence bounds shrink as the arms receive increasing effort from the agents. Once the anchor rate rises above the upper confidence bound of an arm, the policy drops the arm. We are able to prove that this policy generates the optimal amount of experimentation asymptotically even in our conservative case where all agents are fully myopic.

Perhaps surprisingly, if no two arms have the same expected reward rate then a central planner can achieve the optimal expected regret by dropping arms that are suboptimal with probability at least $1 - c_1/T$ if there are at least $c_2 \log T$ agents, for fixed constants c_1 and c_2 which are independent of the time horizon and the number of arms. It turns out that the aggregate expected experimentation effort is no more than $O(J \log T)$. We also observe that the required number of agents does not depend on the number of arms. This is because, with high probability, a lot of arms are discarded fairly quickly because there is enough evidence that they are suboptimal. In fact, we find that after time $\Theta(J)$, thus constant in T, all agents spend all of their effort on the arm with the highest expected reward rate with high probability. Thus, the central planner's policy of dropping suboptimal arms totally distributes the experimentation across the agents and speeds up learning. Prior work on learning with multiple agents shows that with I agents, the learning speed increases by a factor of $\sqrt{I}$ if the agents share information once (Hillel et al. [2013]). The central planner's policy of dropping arms speeds up learning by a factor of I, suggesting that it is a powerful policy.

However, if the arms are hard to separate (such as all but the best arm have the same expected reward rate), then a logarithmic regret cannot be achieved if the number of agents is less than $J \log T$, where J is the number of arms. Note the contrast to the well-separated case where $\Theta(\log T)$ agents is sufficient for optimal exploration irrespective of J. This is because, in this case, there is a very high chance that there is never enough evidence to drop any of the arms and thus no arms are ever discarded. In the centralized version of the problem, the separation between the top two arms is used as a constant to obtain a logarithmic regret. However, in the case of decentralized experimentation, more differentiation among arms is required unless the central planner has access to a lot more agents. This is a cost of decentralization. In the case of the centralized experimentation, only the top two arms need to be different in order to obtain the optimal regret. This condition is insufficient in the decentralized case. Instead, we require that all arms are separated from each other (except in the case where we have at least $J \log T$ agents).

2. Related Work

Multi-armed bandits have long been the canonical framework for studying the problem of experimentation (Whittle [1980]). The particular framework we build on, the finite horizon model with the objective to minimize regret, traces back to Lai and Robbins [1985] and Auer et al. [2002], with Bubeck et al. [2012] serving as a great survey. Classical multi-armed bandits serve as a central-

ized benchmark for us. Several algorithms have been proposed that achieve approximately optimal regret in the centralized setting. The seminal work of Lai and Robbins [1985] showed that any optimal policy must sample each arm at least $\Theta(\log T)$ times to achieve the optimal regret in finite time. Later algorithms showed that such regret is maintained at all times if all arms are pulled for a time equal to the logarithm of the passed time (Auer et al. [2002]). This result was proven via an upper confidence bound (UCB) argument, and this approach motivates the policy we propose.

Recent work on promoting experimentation among agents in decentralized settings has mostly focused on the design of monetary incentive structures and contracts (Halac et al. [2016], Frazier et al. [2014], Chen et al. [2018]) and information design (Kremer et al. [2014], Mansour et al. [2020], Che and Hörner [2018]). For a detailed survey of this literature, please see Chapter 11 in Slivkins et al. [2019]. In Halac et al. [2016], the authors study the design of optimal payment contracts when a principal faces agents that are less interested in experimentation and have private information about their ability to experiment. Frazier et al. [2014] study the tradeoffs between the time-discounted incentive payments made by the principal to myopic agents, and the time-discounted rewards obtained by the principal. They characterize the set of feasible payment and reward pairs and quantify the limitations of payment incentives. In Chen et al. [2018], the authors include heterogeneity among the agents with respect to the reward distributions when pulling arms. This heterogeneity generates natural experimentation and reduces the incentives requires to promote experimentation. An earlier literature in operations management studies learning and inventory planning through sales agents, exploring compensation plans and incentives for such agents (Chen [2000], Chen [2005]). The authors in Chen [2005] study incentives for the salesforce to reveal truthful information about the market and work hard, while in Chen [2000] they study incentives to make demand smoother. While monetary incentives are a powerful tool in certain applications, they are not a practical solution in other settings such as in medical experiments or when learning from user reviews. Our work is a bit more closely related to the policies that use either recommendation or information design as controls for promoting experimentation. Recent work on recommendations to persuade agents to experiment include Bayesian incentive compatible (BIC) exploration (Kremer et al. [2014], Che and Hörner [2018], Papanastasiou et al. [2018], Mansour et al. [2020], Acemoglu et al. [2022]), which in turn builds on the Bayesian persuasion paradigm introduced by Kamenica and Gentzkow [2011].

The most closely related paper to ours is probably Immorlica et al. [2020], a paper that uses selective information disclosure policies to promote experimentation by myopic short-lived users who live for exactly one unit of time. Our proposed policy of adaptively discarding arms can also be considered as an information disclosure policy where the central planner informs everyone that the discarded arm is suboptimal. Immorlica et al. [2020] use sophisticated time varying partitions of the set of agents into groups consisting of disclosure paths such that new agents observe all history of the paths they belong to. The paper explores multiple levels of partitioning where the

disclosure paths from lower levels are connected to paths in the higher levels such that the agents in the higher levels observe the histories of the disclosure paths they are connected to. With a novel and fairly complex technique of merging and interleaving of paths and groups over time, the central planner is able to achieve rate optimal regret as the number of levels increase. In their result, which is agnostic to the separation between arms (see Theorem 6.1 in Immorlica et al. [2020]), they are able to achieve the rate optimal regret up to a polylogarithmic factor using $2^{20}\log^2 T$ groups each consisting of at least $\Omega(T^{1/2})$ disclosure paths in the first level. When adding a minimum separation between arms, the number of needed disclosure paths is smaller (Theorem 6.2 in Immorlica et al. [2020]): they show that with $2^{20}\log^2 T$ groups each with $2^{40}\log^4 T$ disclosure paths in the first level, rate optimal regret upto a polylogarithmic factor can be achieved. Our paper also obtains the rate optimal regret in a fairly different model. Our model does not use short-lived agents (it has myopic long-lived agents instead) and each long-lived agent can be considered to be belonging to their own static singleton group. Thus our disclosure paths are fixed and each group consists of exactly one disclosure path such that the long long-lived agent in the group observes all history of the group. We are able to obtain this regret with a simple policy (dropping arms) and a simple group structure, assuming the arms are separated. Our proposed policy of dropping arms adaptively is sophisticated but easy to implement. The policy we propose is powerful and the number of disclosure paths we need is much smaller, $12\log T$. A critical difference between our approaches is that unlike Immorlica et al. [2020], the central planner does not need groups to settle on arms before making discarding decisions. The diversity in the sample paths across the groups allows for faster information acquisition for the central planner and it can make fast discarding decisions while retaining the information about available arms for future decisions. Thus the experimentation is achieved by the groups who are never allowed to settle on an arm, leading to fewer groups and faster learning. Due to the substantially different nature of our model and the distinct policy, our analysis is both intricate and novel.

On the technical side, our results are also related to the probably approximately correct (PAC) bounds and the sample complexity work in the multi-armed bandits literature (Even-Dar et al. [2002], Mannor and Tsitsiklis [2004]). In finite time, Even-Dar et al. [2006] studied the policies of dropping arms using PAC bounds for multi-armed bandit problems. The PAC bounds are in general closely related to regret bounds because the probability of selecting the optimal arm often determines the regret in finite horizon. However, most of this literature on multi-armed bandits is for actions and rewards in discrete time and is focused on policies that control the sampling of the arms. We study a multi-armed bandit model in continuous time, assuming any pulled arm generates rewards following a Brownian motion with some drift, which is less commonly studied that discrete time models (Mandelbaum [1987], Mandelbaum and Vanderbei [1994], Slivkins and Upfal [2008]). Mandelbaum [1987] showed the existence of well-defined follow-the-leader policies in continuous time and Mandelbaum and Vanderbei [1994] showed the existence of a Gittins index

policy in continuous time and showed it to be optimal. Our work gives a novel characterization of follow-the-leader policies in continuous time for the regret minimization setting when the agents are myopic.

3. Model

We consider a continuous-time learning problem faced by a central planner with $J \geq 1$ arms over a time horizon of length $T \geq 0$. The central planner does not pull the arms themselves but instead offers them to a finite set of $I \geq 1$ agents. The set of arms is denoted by $\mathbb{J} = \{1, 2, ..., J\}$ and the set of agents is given by $\mathbb{I} = \{1, 2, ..., I\}$. The continuous time framework of our model is closest to Mandelbaum [1987] but with the additional feature of a central planner. All of our random variables and processes are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The agents do not communicate and act independently of one another. Moreover, they are assumed to be myopic by considering only immediate rewards while ignoring the future. The central planner may however exert control over them by deciding which arms to make available at each point in time $t \in [0, T]$. Specifically, the set of arms that the central planner makes available at any time t is denoted by $\mathbb{J}_t \subseteq \mathbb{J}$. Moreover, we impose the constraint that $\mathbb{J}_s \supseteq \mathbb{J}_t$ for each $s < t$. That is, once the central planner removes an arm from the agents, it cannot be added back later. We also require that $\mathbb{J}_T \neq \emptyset$. That is, the central planner cannot remove all of the arms.

One valid question is why do we assume the agents are myopic given they are long-lived. For the problem to be interesting, the agents need to be less interested in experimentation than the central planner. We could have modeled this discrepancy by assuming different discount rates. There are two reasons we did not make this choice. First, this would complicate the model by forcing us to replace the agent's static decision-making problem by a complex dynamic one. Second, and more importantly, this would likely make learning easier. By assuming agents are myopic, we are making the most conservative assumption possible regarding how much exploration agents deliberately add to the system: none. Still, we are able to obtain rate-optimal learning results despite this conservative assumption in the presence of relatively few agents. Making the agents less myopic should, in principle, make learning easier and allow the central planner to learn with even fewer agents.

For each agent $i \in \mathbb{I}$, we denote their cumulative effort allocation vector to each arm at time $t \geq 0$ by the J-dimensional vector $\boldsymbol{\tau}^i(t) = (\tau_1^i(t), \tau_2^i(t), ..., \tau_J^i(t))$. We assume that each τ_j^i is non-decreasing with $\tau_j^i(0) = 0$. Moreover, we require that $\tau_1^i(t) + ... + \tau_J^i(t) = t$. This implies that at each point in time the agent must exert full effort across the arms. The agent can however divide their effort arbitrarily across different arms. There is however one exception. An agent cannot pull an arm which has already been removed by the central planner. This condition is enforced by

requiring that for each arm $j \in \mathbb{J}$ and time $t \geq 0$,

$$\tau_j^i(t) \quad = \quad \int_0^t 1\{j \in \mathbb{J}_s\} d\tau_j^i(s). \tag{1}$$

Now suppose that over the interval of time $[t, t+\Delta]$, agent $i \in \mathbb{I}$ pulls arm $j \in \mathbb{J}$ with effort $\Delta\tau_j^i(t)$. We then assume that the cumulative reward received by agent i from arm j over the interval of time $[t, t+\Delta]$ is given by the normal random variable $N(\mu_j \Delta\tau_j^i(t), \Delta\tau_j^i(t))$. Moreover, rewards received over disjoint intervals of time are independent of one another. In this case, agent i's cumulative reward up until time t from pulling arm j is given by

$$R_j^i(\tau_j^i(t)) = \mu_j \tau_j^i(t) + B_j^i(\tau_j^i(t)),$$

where B_j^i is a standard Brownian motion. Neither the central planner nor the agents know the vector $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_J)$. They do however know the general form of the reward function. Moreover, we assume that the family $\{B_j^i(\cdot), i \in \mathbb{I}, j \in \mathbb{J}\}$ of standard Brownian motions are independent of one another.

The quantity μ_j in the above is the expected instantneous reward rate of the agent at each point in time assuming it dedicates all of its efforts into pulling arm j. We henceforth refer to μ_j as the drift term of arm j. Note that by assumption the drift of arm j is the same across all of the agents. On the other hand, each agent i has their own set $\{B_j^i(\cdot), j \in \mathbb{J}\}$ of Brownian motions for each arm. The practical interpretation is that all agents are equally skilled and thus have the same expected reward rate and variance from a given arm, however each agent encounters their own idiosyncratic source of randomness when pulling it.

Now note that if each agent were clairvoyant and knew the drift vector $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_J)$, then in order to maximize their expected cumulative reward up until time T, each agent would put all of its effort into pulling the arm with the highest drift. Instead, since each agent is not clairvoyant, they may instead continually construct and update estimates of the drift of each arm. However, since by assumption agents do not share information, they may only base their estimate on their own experience with each arm. In particular, we assume that agent i's time t drift estimate of arm j assuming $\tau_j^i(t) = s$ is given by

$$\hat{\mu}_j^i(s) = \frac{1}{1+s} R_j^i(s) = \frac{1}{1+s} \left(\mu_j s + B_j^i(s)\right). \tag{2}$$

Note in particular that at each point in time agents will have different drift estimates of the same arm. This is because of their own idiosyncratic noise and the fact that agents may have allocated different amounts of effort to the arm.

There are two ways to interpret this drift estimate. First, the above may be thought of as a regularized frequentist estimate of the drift. The unregularized frequentist estimate is $R_j^i(s)/s$

but such an estimate is ill-defined at $s = 0$. The addition of 1 (or any positive constant) to the denominator ensures the estimator is well-behaved near effort zero while not affecting its asymptotic properties. This estimator may also be thought of as a Bayesian estimate of the drift μ_j. To do so, one would assume all agents' prior beliefs on the distribution of μ_j are Gaussian with mean zero and variance one. Note also that $\hat{\mu}_j^i(s)$ is not the estimate of the drift at time s, but instead the estimate at any time t such that effort is equal to s, i.e. $\tau_j^i(t) = s$.

Next, recall that the agents are myopic and at each point in time they choose an action that optimizes their expected instantaneous reward. Such a strategy is referred to in Mandelbaum [1987] as a follow the leader strategy. It is defined by requiring that for each $t \geq 0$ and $j \in \mathbb{J}_t$, τ_j^i increases at time t only if arm j has the highest drift estimate at time t among all remaining arms. Setting

$$k^i(t) = \arg\max_{m \in \mathbb{J}_t} \hat{\mu}_m^i(\tau_m^i(t)), \tag{3}$$

the myopic policy is technically defined by the requirement that $\tau_j^i(u) > \tau_j^i(t)$ only if $j \in k(t)$ for all $u > t$. Note also this implies that $\tau_j^i(t)$ cannot increase if arm j has been removed by time t, that is $j \notin \mathbb{J}_t$. It follows in a straightforward manner from Propositions 2 and 5 of Mandelbaum [1987] that the myopic policy $\boldsymbol{\tau}^i = (\tau_1^i, \tau_2^i, ..., \tau_J^i)$ thus described is unique.

The central planner observes all of the efforts and rewards obtained by the agents. Thus, the information available to the central planner at time $t > 0$ is given by the σ-algebra

$$\mathcal{H}_t \quad = \quad \sigma(\tau_j^i(s), R_j^i(\tau_j^i(s))) \text{ for all } s \in [0, t) \ , i \in \mathbb{I}, j \in \mathbb{J}.$$

A policy $\pi = \{\pi_t, t \in [0, T]\}$ of the central planner is an $\mathcal{H}_t$-adapted process such that $\pi_t \subset \mathbb{J}$ for each $t \in [0, T]$ and $\pi_s \subseteq \pi_t$ for $s < t$. The set π_t represents the set of all arms that the central planner has chosen to remove from its lineup up until time t. Thus, the set of arms available at time t is given by $\mathbb{J}_t = \mathbb{J} \setminus \pi_t$. A policy π is feasible if π_t is a right-continuous function and $\pi_T \neq \mathbb{J}$. This ensures that the time at which an arm is removed from the lineup is well-defined and that there is always at least one arm available for the agents to pull. The set of all feasible policies is denoted by Π.

Given a policy $\pi \in \Pi$, the central planner's utility U_π is equal to the sum of the agents' accumulated rewards up until time T. Specifically,

$$U_\pi(T, \boldsymbol{\mu}) = \sum_{i=1}^{I} \sum_{j=1}^{J} R_j^i(\tau_j^i(T)).$$

Note that π and $\boldsymbol{\mu}$ appear on the lefthand but not the righthand side above. They do however as a consequence of the discussion above control the reward processes on the righthand side.

Now note that if the central planner was clairvoyant and knew the drift vector $\boldsymbol{\mu}$, it would

at the outset discard all of the arms except the one with the highest drift. The expected utility thus earned would be $IT \max_{j\in\mathbb{J}} \mu_j$. Given a policy π, the central planner's regret is defined to be the difference between the expected utility it would obtain if it knew the drift vector $\boldsymbol{\mu}$ and the expected utility it obtains under the chosen policy π. That is, the central planner's regret is given by

$$\tilde{Z}_\pi(T, \boldsymbol{\mu}) = IT \max_{j\in\mathbb{J}} \mu_j - \mathbb{E}[U_\pi(T, \boldsymbol{\mu})]. \tag{4}$$

The central planner's worst-case regret with respect to the drift vector $\boldsymbol{\mu}$ is denoted by

$$Z_\pi(T) = \sup_{\boldsymbol{\mu}\in\mathbb{R}^J} \tilde{Z}_\pi(T, \boldsymbol{\mu}).$$

The objective of the central planner is to find a policy π that minimizes its worst-case regret with respect to the drift vector, i.e. $Z^* = \inf_{\pi\in\Pi} Z_\pi(T)$. We focus on finding a policy $\pi \in \Pi$ for which we can provide strong performance guarantees, in the sense of a rate-optimal regret.

4. Single Agent with No Central Planner

In this section, we consider the scenario when a central planner does not exist. The agents are therefore free to allocate their effort to any arm in the set $\mathbb{J}$ all the way up to the final time T. We do however still assume that the agents are myopic and operate according to the follow the leader policy described in Section 3. Another way to think about this case is that the central planner does exist but sets $\mathbb{J}_t = \mathbb{J}$ for all $t \in [0, T]$.

One reason for studying this case first is that it provides us with a lower bound on the expected reward against which to compare the improvement with a central planner. It also turns out that in this case we are able to obtain sharp results and the methodology we develop in this section will be used to prove the results in later sections where multiple agents and a central planner are involved.

Recall now that in the model given in Section 3 the agents do not communicate with one another. It then follows that without a central planner, each of the agents follows a myopic policy independent of one another. We therefore, for the sake of simplicity, assume in this section without loss of generality that there exists a single agent. We also drop the superscript i from our notation throughout this section.

4.1 Structural Results

In this subsection, we establish some structural properties of the myopic policy from Section 3 when followed by a single agent in the absence of a central planner. We begin by defining for each arm

$j \in \mathbb{J}$ and any time t with cumulative effort $\tau_j(t) = t^\circ \in [0, T]$ the quantity

$$\underline{\mu}_j(t^\circ) = \min_{0 \leq s \leq t^\circ} \hat{\mu}_j(s), \tag{5}$$

which is referred to as the lower envelope of $\hat{\mu}_j$. It turns out (see the discussion in section 4.4 of Mandelbaum [1987]) that the unique myopic policy defined in Section 2 is the same as the unique myopic policy which at each point in time pulls the arm with the largest lower envelope. Thus, for each arm $j \in \mathbb{J}$ and time $t \in [0, T]$, let

$$L_j(t) = \underline{\mu}_j(\tau_j(t)) = \min_{0 \leq s \leq t} \hat{\mu}_j(\tau_j(s)) \tag{6}$$

denote the agent's minimum estimate of the drift of arm j up until time t. It is then easily shown to be a consequence of Proposition 1 of Mandelbaum [1987] that at any given point in time the running minimum estimates of all the available arms are almost surely identical. Specifically, we have the following.

Lemma 1. *For each $t \in [0, T]$ and $j, l \in \mathbb{J}_t$, $L_j(t) = L_l(t)$.*

For completeness, we also provide a short proof of the Lemma in Appendix. We note that by definition $L_j(t)$ is non increasing for each arm j. Following the Lemma, we henceforth use the shorthand $L(t)$ to refer to the agent's running minimum estimate of all of the available arms at time t. The following is then a consequence of Proposition 5 of Mandelbaum [1987]. It implies that at almost every point in time the myopic agent will devote all of their effort to a single arm.

Lemma 2. *In the absence of a central planner, for each $t \in [0, T]$ the set of arms $k(t)$ with the highest drift estimates is a singleton with probability 1.*

We next provide additional results on the system dynamics to better characterize the arm with the highest drift estimate at each point in time. These results will also help to characterize the cumulative effort on each arm up until each point in time. It turns out that the collection $\{L_j(t), j \in \mathbb{J}\}$ of running minimum estimates are crucial for understanding the behavior of the system. Furthermore, at almost all times the current estimate of each arm is equal to its running minimum estimate, except for the estimate of a single arm corresponding to the highest current estimate. Finally, the minimum estimate of the drift of each arm, other than the arm with the highest current esimtate, would strictly increase (decrease) with any strictly lower (higher) effort allocated to it. We have the following.

Theorem 1. *For each $t \in [0, T]$, the following statements hold almost surely:*

(i) If $j \in k(t)$, then $\hat{\mu}_j(\tau_j(t)) > L_j(t)$ and if $j \in \mathbb{J}_t \setminus k(t)$, then $\hat{\mu}_j(\tau_j(t)) = L_j(t)$.

(ii) If $j \in \mathbb{J}_t \setminus k(t)$, then $\hat{\mu}_j(y) > L_j(t)$ for all $y < \tau_j(t)$ and $\underline{\mu}_j(y) < L_j(t)$ for all $y > \tau_j(t)$.

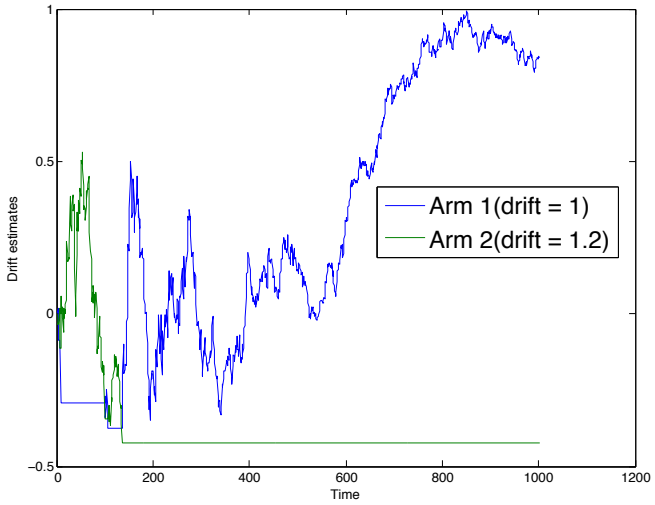

Figure 1: Drift estimates of two available arms for one agent as a function of time. We point that the x-axis represents the clock time and not the cumulative effort on the arms. The drift estimate of a arm does not change if it does not receive any effort.

We now demonstrate the above Lemmas and Theorem 1 through the following example. Consider two available arms with drifts of 1 and 1.2. Figure 1 depicts a sample path of the agent's drift estimates for the two arms. Note that the drift estimate of only one arm changes at any time. This is because the drift estimates of the two arms are different at almost all times and the agent applies all its effort on the arm with the higher drift estimate. This is expected from Lemma 2. However, also note as a consequence of Lemma 1 that the running minimum estimates of the two arms are at all times equal to one another.

Another way to demonstrate this is by observing how the estimates of arm 1 changes relative to the drift estimate of arm 2. We show this in Figure 2. Note in the figure that the running minimum estimates of both arms until the current time are equal and the lower of the current drift estimates of the two arms is equal to the running minimum estimates at the current time. We also observe that at any time the arm that the agent puts effort on has a drift estimate higher than the running minimum estimates and the other arm's drift estimate are equal to the running minimum estimates as suggested by Theorem 1.

An important implication of Theorem 1, as also observed in Figure 1, is that any time t the cumulative effort $\tau_j(t)$ for any arm $j \neq k(t)$ is the minimum effort for which the minimum drift estimate of the arm is equal to the common minimum drift estimate of all arms. We define the minimum effort required for the drift estimate for an arm j to achieve a given quantity $x \leq 0$ as

$$\sigma_j(x) = \inf\{s > 0 : \hat{\mu}_j(s) \leq x\},$$

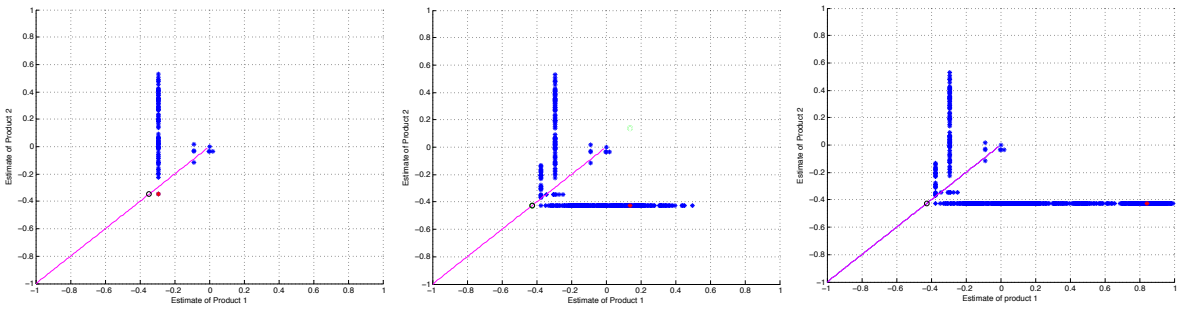

Figure 2: Drift estimates of the two arms plotted against each other. The first figure captures the drift estimates until time 100, the second figure until time 500 and the third figure until time 1000. The x-axis in all figures is the drift estimate of arm 1 and the y-axis represents the drift estimate of arm 2. The red '*' represents the current estimates. The black 'o' represents the running minimum estimates of both arms at the current time as well as the lower of the current estimates of the two arms. We point out that the gaps are due to the discontinuity introduced by the simulation.

where $\inf \emptyset = \infty$. The random variable $\sigma_j(x)$ therefore takes values in the extended real line $\mathbb{R}_+ \cup \infty$. We now state the following immediate corollary of Theorem 1 that connects the cumulative effort on any arm $j \neq k(t)$ to the common minimum drift estimate of all arms through σ_j.

Corollary 1. *For each arm $j \neq k(t)$, $\tau_j(t) = \sigma_j(L(t))$ almost surely.*

This result provides us a way to characterize the distribution of efforts on arms through the distribution of the common minimum drift estimate of all the arms until the current time. We will later in this section characterize this distribution and use it for analyzing the regret.

4.2 The Case of a Single Arm

We first consider the case of a single arm. We will drop the subscript j in this subsection because it is understood that there is only one arm. The dynamics of a single agent with a single arm are the same whether there is a central planner or not since by assumption the central planner must always keep at least one arm. We proceed as follows.

First note that since there exists only one arm, the agent will at each point in time dedicate their full effort to it. That is, $\tau(t) = t$ for $t \in [0, T]$. It then follows by (2) that the agent's time t drift estimate of the arm is given by

$$\hat{\mu}(\tau(t)) \quad = \quad \hat{\mu}(t) \quad = \quad \frac{1}{1+t}(\mu t + B(t)), \tag{7}$$

where B is a standard Brownian motion. Using the fact that $B(t)$ is normally distributed with a mean of 0 and a variance of t, one may easily use the above to compute distribution of the agent's drift estimate at each point in time.

Now recall from Section 4.1 the definition for each t of $L(t)$ as the minimum running drift estimate of all of the arms by time t. This quantity in the current case is simply equal to the agent's minimum drift estimate of the single arm up until time t. That is,

$$L(t) = \min_{0 \leq s \leq t} \hat{\mu}(\tau(s)). \tag{8}$$

The quantity on the righthand side above at first glance appears to have a difficult to compute distribution function. It turns out however that it is easier to work with the inverse function of L instead. We proceed as follows.

First note that the running minimum function L is continuous and non-increasing. We may therefore define its inverse function

$$L^{-1}(x) = \inf\{t \geq 0 | L(t) \leq x\}$$

for $x \leq 0$, where $\inf \emptyset = \infty$. In other words, if the agent's estimate is always above the level x, then $L^{-1}(x) = \infty$. The random variable $L^{-1}(x)$ therefore takes values in the extended real line $\mathbb{R}_+ \cup \infty$.

We now point out that for the case of the single arm, $L^{-1}(x) = \sigma(x)$ because when there is only one arm then $\inf\{t \geq 0 | L(t) \leq x\} = \inf\{t \geq 0 | \hat{\mu}(t) \leq x\}$. It then follows that for each $t \geq 0$, and $x \leq 0$ we have $L(t) \leq x$ if and only if $L^{-1}(x) \leq t$. Consequently, taking expectations $\mathbb{P}(L(t) \leq x) = \mathbb{P}(L^{-1}(x) \leq t)$. On the other hand, by (7) and (8) it follows after some algebra that

$$\sigma(x) = L^{-1}(x) = \inf\{t \geq 0 | B(t) = x + (x - \mu)t\}. \tag{9}$$

We therefore see that $L^{-1}(x)$ may be characterized as the first hitting time of an affine barrier with intercept x and slope $(x - \mu)$ by a standard Brownian motion. The distribution of this random variable is well-known and so we obtain the following result.

Lemma 3. *In the case of a single agent with a single arm, for each $t \geq 0$ and $x \leq 0$,*

$$\mathbb{P}(L(t) \leq x) = \mathbb{P}(\sigma(x) \leq t) = e^{2x(\mu - x)} \Phi\left(\frac{(\mu - x)\, t + x}{\sqrt{t}}\right) + 1 - \Phi\left(\frac{(\mu - x)\, t - x}{\sqrt{t}}\right), \tag{10}$$

where Φ is the c.d.f. of the standard normal distribution.

The exact expression on the right hand side above for the distribution of the time t minimum drift estimate does not carry over to the general case of a central planner with multiple agents. It is still however useful in proving our main results by serving as the starting point for deriving comparative results for the hitting times of arms with different drifts. In particular, we have the following result as a consequence of the above Lemma.

Lemma 4. *In the case of a single agent with a single arm, for each $t > 0$, the minimum drift estimate $L(t)$ is stochastically increasing in a first order sense with respect to the drift μ. Similarly,*

for each $x < 0$, the hitting time $L^{-1}(x)$ or $\sigma(x)$ is stochastically increasing in a first order sense with respect to the drift μ.

The above Lemma is useful for comparing distributions of effort allocation for multiple arms and consequently the regret when there are multiple arms. We discuss the case of multiple arms next.

4.3 The Case of Multiple Arms

We now proceed to the case of a single agent with multiple arms ($J \geq 1$) but still with no central planner. The drifts in this case for each arm are given by the drift vector $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_J)$. Moreover, the agent's time $t \in [0, T]$ drift estimate for each arm $j \in \mathbb{J}$ is given by $\hat{\mu}_j(\tau_j(t))$, where $\tau_j(t)$ is the cumulative effort that the agent has allocated to arm j up until time t, and $\hat{\mu}_j(\cdot)$ is as given in equation (2).

The interesting feature in this case is that the agent may spread their effort out across each of the arms in a somewhat complicated way. We therefore do not have a simple expression for the effort allocation vector $\boldsymbol{\tau}(t) = (\tau_1(t), \tau_2(t), ..., \tau_J(t))$. On the other hand, recall from Section 4.1 the definition

$$L_j(t) = \min_{0 \leq s \leq t} \hat{\mu}_j(\tau_j(s)), \tag{11}$$

for each arm $j \in \mathbb{J}$ of the minimum drift estimate up until time t. Then, by Lemma 1 it follows that the minimum drift estimate of all arms are equal, i.e. $L_1(t) = L_2(t) = ... = L_J(t)$ at each time t. We henceforth denote the common minimum drift estimate at time t by $L(t)$. Now for each arm $j \in \mathbb{J} \setminus k(t)$ we recall from Corollary 1 that for all arms $j \neq k(t)$, $\tau_j(t) = \sigma_j(L(t))$ and for $j = k(t)$ $\tau_j(t) \geq \sigma_j(L(t))$. This relationship between the minimum estimates of the drift of each arm is used to derive the results below.

Lemma 5. *In the case of a single agent with $J \geq 1$ arms and no central planner, for each $x, t \geq 0$, $L(t) \leq x$ if and only if $\sigma_1(x) + \sigma_2(x) + ... + \sigma_J(x) \leq t$ almost surely.*

Lemma 4 characterizes in the single arm case the impact of the arm's drift on the effort required to reach a given value of the running minimum drift estimate. This following is a generalization of Lemma 4 to the case of multiple arms.

Lemma 6. *In the case of a single agent with $J \geq 1$ arms and no central planner, for each $t > 0$, the minimum drift estimate $L(t)$ is stochastically increasing in a first order sense with respect to the drift μ_J. Similarly, for each $x < 0$, the sum $\sigma_1(x) + \sigma_2(x) + ... + \sigma_J(x)$ is stochastically increasing in a first order sense with respect to the drift μ_J.*

The above lemma continues to hold if instead of varying the drift of arm J, we vary the drift of any of the $j = 1, 2, ..., J$ arms. This result generalizes Lemma 4 and allows a comparison between

effort allocations for different subsets of arms given the two sets can be ordered. We also point out that the powerset of the set of arms may not have a linear order with respect to stochastic dominance. The exact distribution of the cumulative effort allocation for any subset of arms can be obtained by a convolution operation and using equation 10. However, the above result is sufficient for our analysis. Using the above result, we next quantify the asymptotic regret for a single agent and multiple arms, in the absence of a central planner.

4.4 Regret Analysis

We now analyze the regret of the expected reward in the case of a single myopic agent in the absence of a central planner relative to the expected reward of a clairvoyant agent in the absence of a central planner. It is clear that the clairvoyant agent will pick the arm with the highest drift at time 0 and then stay with it until the end of the time horizon. Denoting by $\mu_{max} = \max_{j \in \mathbb{J}} \mu_j$ the highest drift amongst all of the arms, the expected reward of the clairvoyant agent is then given by $T\mu_{max}$.

We now analyze the reward received by the myopic agent. First note that due to the inherent randomness of the reward processes, there is no need to force the agent to put non-zero effort into each arm. This is in contrast to the practice in the discrete time multiarmed bandit setting. Instead, in the continuous time setting a myopic agent will almost surely put positive effort into each arm. Moreover, the agent will eventually settle on one arm and then never switch again.

The arm that the agent eventually settles on may be explicitly identified. For each arm $j \in \mathbb{J}$, let

$$M_j(\infty) = \min_{s \geq 0} \hat{\mu}_j(s) \tag{12}$$

denote the all-time minimum of the agent's drift estimate for it, assuming the agent puts infinite effort into pulling the arm. A straightforward argument using the results of the previous section then shows the arm that the agent eventually settles on is given by

$$k^\star \quad = \quad \arg\max_{j \in \mathbb{J}} M_j(\infty). \tag{13}$$

We first point out two important properties of the all-time minimum drift estimates. The first property is that it they each have a finite expectation and the second is that the expected effort to attain the all-time minimum drift estimate is finite. We state these properties formally in the following theorem.

Theorem 2. *For any $\mu_j \in \mathbb{R}$, the expected value of the all-time minimum drift estimate and the expected effort needed to attain the all-time minimum drift estimate are finite constants that depends upon μ_j i.e. $E[M_j(\infty)] = c_1(\mu_j) > -\infty$ and $E[\sigma_j(M_j(\infty))] = E[L_j^{-1}(M_j(\infty))] = c_2(\mu_j) < \infty$ for all $\mu_j \in \mathbb{R}$, where $c_1(\mu_j)$ and $c_2(\mu_j)$ are finite constants that depend upon μ_j.*

The theorem follows from Lemma 3. Specifically, by Lemma 3 the distributions of the all-time minimum drift estimate as well as the hitting time of the all-time minimum drift estimate have exponentially decaying tails. This implies that they have finite expectations. An important implication of the Theorem along with Theorem 1 is that it provides a constant upper bound (independent of the time horizon) on the expected amount of effort the agent spends on all arms other than the one it settles on. This suggests that the amount of experimentation a myopic agent generates is very small and is upper bounded by a constant amount, irrespective of the time horizon.

We are now in a position to provide an asymptotic result on the expected regret of a single myopic agent in the absence of a central planner. To see this, first note that on any particular sample path the myopic agent will eventually settle on arm $k^\star$ and so their realized reward for that sample path will be $\mu_{k^\star}T + \Theta(J)$ as $T \to \infty$. Following the definition of regret for a given $\boldsymbol{\mu}$,

$$\tilde{Z}(T, \boldsymbol{\mu}) = T\mu_{max} - \mathbb{E}\left[\sum_{j\in\mathbb{J}} R_j(\tau_j(T))\right] \tag{14}$$

is the expected regret for the single agent. Then, since other than a finite expected effort bounded by a constant, independent of the time horizon, T, the myopic agent spends all its time on the arm it settles on, we arrive at the following result.

Theorem 3. *For each $J \geq 1$ and $\boldsymbol{\mu} \in \mathbb{R}^J$, it follows that*

$$\tilde{Z}(T, \boldsymbol{\mu}) \;=\; T\sum_{j\in\mathbb{J}}(\mu_{max} - \mu_j)\mathbb{P}k^\star = j|\boldsymbol{\mu}) + \Theta(J) \text{ as } T \to \infty. \tag{15}$$

We are able to obtain an expression for the probability $\mathbb{P}(k^\star = j|\boldsymbol{\mu})$ that the myopic agent settles on an arm j. We provide it in the appendix A.1.3. For general $\boldsymbol{\mu}$, this expression is represented as a sum. However, when all arms have positive drift, then the expression for the probability is much simpler. We provide the expression in the following theorem.

Theorem 4. *Assume there are J arms and the drifts of all arms are positive. The probability that the winning arm is j,*

$$\mathbb{P}\left(k^\star = j|\boldsymbol{\mu}\right) = \frac{1}{J} + \frac{\sqrt{2\pi}\left(\mu_j - \bar{\mu}\right)}{\sqrt{J}} e^{\frac{J\bar{\mu}^2}{2}} \Phi\left(-\sqrt{J}\bar{\mu}\right),$$

where $\bar{\mu} = \frac{1}{J}\sum_{k\in\mathbb{J}}\mu_k$ and Φ is the cdf of the standard normal random variable.

The above theorem provides a simple expression for the probability. In particular the favorability of an arm to end up as the arm the agent settles on is proportional to the difference between its drift and the average drift of all arms. An arm with above average drift will have positive favorability and an arm with below average drift will have negative favorability. Thus, we find that the regret of a myopic agent grows linearly with time and we are able to obtain the linear coefficient as well.

5. Single Agent with a Central Planner

We now study the scenario with a single agent and a central planner. This will help to quantify the impact of the central planner on the regret of a myopic agent. Note that for any given policy π of the central planner, many of the agent related quantities from the previous sections are still well-defined. Specifically, given a policy π, we denote by $L_\pi(t)$ the minimum running drift estimate of the myopic agent at time t. Moreover, for each arm $j \in \mathbb{J}$ we denote by $\tau_{\pi,j}(t)$ the cumulative effort put into arm j by time t. Also, $k_\pi(t)$ is the set of arms with the highest drift estimates at time t. Expressions for $L_\pi(t), \tau_{\pi,j}(t)$ and $k_\pi(t)$ may be written in terms of the underlying Brownian motions however they are rather complicated. Nevertheless, we may still prove useful results for these quantities as we show below. We first note that Lemmas 1 and 2 as well as Theorem 1 still apply in the presence of the central planner because those results only consider the set of available arms at any time. Further, because the lower envelope of the drift estimate is non-increasing in time, the following corollary of Lemma 1 holds.

Corollary 2. *Under any policy π, for each $t \in [0,T]$ and each pair of arms $j \in \mathbb{J}_t$ and $m \notin \mathbb{J}_t$, $L_{\pi,m}(t) \geq L_{\pi,j}(t)$ almost surely.*

To study the quantities $L_\pi(t), \tau_{\pi,j}(t)$ and $k_\pi(t)$ further we first consider a simple policy. Let $K \subset \mathbb{J}$ be a subset of arms and define the policy π_K by setting $\pi_K(t) = \mathbb{J} \setminus K$ for $t \geq 0$. That is, the central planner at time 0 removes all of the arms in the set $\mathbb{J} \setminus K$ and takes no further action to remove additional arms. Thus, over the course of the time horizon only arms in the set K are available to the agent. This seemingly simple policy serves as a building block for more complicated policies of the central planner in which arms are removed at different points in time. The following result characterizes the relationship between an arbitrary policy π and the policy π_K for arms in $K \subseteq \mathbb{J}$.

Lemma 7. *Let π be an arbitrary policy of the central planner and for $K \subseteq \mathbb{J}$ set*

$$t^\circ \;=\; \sum_{l \in K} \tau_{\pi,l}(t). \tag{16}$$

Then, for $\mathbb{P}$-a.e. $\omega \in \Omega$, if $K \subseteq \mathbb{J}_t(\omega)$, the following statements are true:

(i) $L_{\pi_K}(t^\circ) = L_\pi(t)$.

(ii) $j \in k_{\pi_K}(t^\circ)$ *if* $j \in k_\pi(t)$.

(iii) $\tau_{\pi_K,j}(t^\circ) = \tau_{\pi,j}(t)$ *for all* $j \in K$.

The first statement of the lemma connects the minimum running drift estimate of the policy π at time t to the minimum running drift estimate under the policy π_K at time t°. The second

statement says that if an arm $j \in K \subseteq \mathbb{J}_t$ is a winning arm under policy π at time t, then it is also a winning arm under π_K at time t°. The third statement says that the cumulative effort on an arm $j \in K \subseteq \mathbb{J}_t$ under the policy π at time t is the same as the cumulative effort on an arm j under the policy π_K at time t°. This helps us to compare the cumulative effort on two arms available at time t under any policy π as stated in the following result.

Theorem 5. *Let π be an arbitrary policy of the central planner and $K \subseteq \mathbb{J}$. If $j_1, j_2 \in K$ with $\mu_{j_1} > \mu_{j_2}$, then*

$$1\{K \subseteq \mathbb{J}_t\}\mathbb{P}\left(\tau_{\pi,j_1}(t) > s \,\middle|\, \sum_{j\in K}\tau_{\pi,j}(t)\right) \;\geq\; 1\{K \subseteq \mathbb{J}_t\}\mathbb{P}\left(\tau_{\pi,j_2}(t) > s \,\middle|\, \sum_{j\in K}\tau_{\pi,j}(t)\right) \quad (17)$$

for $0 < s < \sum_{j\in K}\tau_{\pi,j}(t)$. Moreover, the inequality above is strict if $K \subseteq \mathbb{J}_t$.

Another important result we need is if arm $j \in K$ is available at time t under the policy π, then it statistically has a higher cumulative effort than under the policy π_K at time t° assuming (16) holds. We note the subtlety that all arms in K need not be available at time t under policy π.

Theorem 6. *Let π be an arbitrary policy of the central planner and $K \subseteq \mathbb{J}$, then*

$$1\{j \in K \cap \mathbb{J}_t\}\mathbb{P}\left(\tau_{\pi,j}(t) > s \,\middle|\, \sum_{l\in K}\tau_{\pi,l}(t) = t^\circ\right) \geq 1\{j \in K \cap \mathbb{J}_t\}\mathbb{P}\left(\tau_{\pi_K,j}(t^\circ) > s\right)$$

for all $0 < s < t^\circ$ and all $t^\circ > 0$. The equality holds only if $K \subseteq \mathbb{J}_t$.

The proofs of Theorems 5 and 6 use two additional lemmas that may be found in the appendix.

We now lower bound the minimal regret assuming one myopic agent and a central planner. Note that even if the central planner is forward looking, it has limited control over the actions of the myopic agent. As Theorem 2 of Section 4.4 suggests, the uncertainty in rewards results in the agent only experimenting a finite amount of time before settling on an arm.

In order to generate more experimentation on arms that are likely to be good arms, the central planner would have to discard arms early enough before the agent settles on an arm. However, this risks discarding good arms. Discarding an arm that the agent eventually settles on can restart the experimentation by forcing the agent to try among the remaining arms but this is even riskier because the arm that the agent settles on is likely the best arm as suggested by Theorem 4. It turns out that in the case of a single agent the best policy of the central planner is to never remove an arm because the risk of removing the best arm exceeds the benefits of generating more experimentation. Therefore, as in the case of a single agent without a central planner, the regret grows linearly in time.

Theorem 7. *For each $J \geq 1$ and $\boldsymbol{\mu} \in \mathbb{R}^J$, it follows that under any policy π of the central planner*

$$\tilde{Z}_{\pi}(T, \boldsymbol{\mu}) \quad \geq \quad T \sum_{j \in \mathbb{J}} (\mu_{max} - \mu_j) \mathbb{P}(k^{\star} = j | \boldsymbol{\mu}) + \Theta(J) \text{ as } T \to \infty. \tag{18}$$

In (18), the probabilities $\mathbb{P}(k^{\star} = j|\boldsymbol{\mu})$ of the myopic agent settling on arm j are the same as in the case of a single agent without a central planner as discussed in Section 4.4. Thus, we have established that the myopic agent's regret grows linearly with time and a central planner cannot improve the regret if there is only one agent. In the next section, we study if the central planner can improve the regret when there are multiple myopic agents.

6. Multiple Agents with a Central Planner

We now consider the case of a central planner with multiple agents. Note that with multiple agents the central planner may aggregate the information from all of the agents and discard arms at a lower risk of discarding the optimal arm than with a single agent. However, the central planner needs to tradeoff the risk of discarding the optimal arm with the risk of agents spending too much effort on suboptimal arms.

In this section, we study an efficient policy of the central planner with multiple agents. We first establish a benchmark on the asymptotics of the minimal regret of the central planner assuming it had full control of the agents. This regret grows logarithmically in time and provides a lower bound on the regret of the central planner if it can only discard arms. We aim to achieve this regret bound using an efficient policy.

In the previous section, we established that if there is only one agent, then any policy of discarding arms generates at best a linear regret. The question then arises of "how many agents are needed to achieve a logarithmic regret?" We first show that if the central planner does not have control over the agents and if there are $o(\log T)$ agents, then no admissible policy of the central planner can asymptotically achieve a logarithmic regret. We then introduce an asymptotically efficient policy for the central planner and show that it achieves the logarithmic benchmark if there are at least $12 \log T$ agents.

Due to the fact that we now have multiple agents, it is necessary to introduce some additional notation. Let π be an arbitrary policy of the central planner. The cumulative effort spent on arm $j \in \mathbb{J}$ until time t by each agent is then denoted by the vector $\boldsymbol{\tau}_{\pi,j}(t) = \left(\tau^1_{\pi,j}(t), \ldots, \tau^I_{\pi,j}(t)\right)$. The total effort spent on arm j until time t by all the agents is

$$\tau^f_{\pi,j}(t) = \mathbf{1}^T \boldsymbol{\tau}_{\pi,j}(t) = \sum_{i=1}^{I} \tau^i_{\pi,j}(t).$$

The superscript f is used to denote the fact that the central planner has full information over all

effort levels and rewards obtained by the agents. Given a cumulative arm j effort vector $\boldsymbol{s}_j$, the total arm j reward accumulated by all the agents is $R_j^f(\boldsymbol{s}_j) = \sum_{i=1}^I R_j^i(s_j^i)$. Hence, given that the central planner uses a policy π, the total reward received by all agents from arm j until time t is

$$R_j^f(\boldsymbol{\tau}_j(t)) = \sum_{i=1}^I R_j^i(\tau_j^i(t)).$$

Now note that for any arm $j \in \mathbb{J}$ and any two effort vectors $\boldsymbol{s}_j, \tilde{\boldsymbol{s}}_j \geq 0$ with $\mathbf{1}^T \boldsymbol{s}_j = \mathbf{1}^T \tilde{\boldsymbol{s}}_j$, the total arm j rewards $R_j^f(\boldsymbol{s}_j)$ and $R_j^f(\tilde{\boldsymbol{s}}_j)$ are statistically equivalent.

For each agent $i \in \mathbb{I}$, we denote by $\boldsymbol{\tau}_\pi^i(t) = \left(\tau_{\pi,1}^i(t), \tau_{\pi,2}^i(t), ..., \tau_{\pi,J}^i(t)\right)$ the cumulative effort spent by agent i on each arm until time t under the policy π.

6.1 Regret Bounds

Lai and Robbins [1985] established that for the case of discrete-time multi-armed bandits the best possible regret over all $\boldsymbol{\mu} \in \mathbb{R}^J$ is of order $\Theta(J \log T)$ given a time horizon T. In this subsection, we extend their analysis to the case of continuous-time bandit processes. The setting of Lai and Robbins [1985] concerns a centralized experimentation problem. In our context, this is equivalent to assuming that $\mathbb{J}_t = \mathbb{J}$ for $t \geq 0$ and there exists a single agent who may select any cumulative effort vector process $\boldsymbol{\tau}$ that is $\mathcal{H}_t$-adapted. In the following result, we use the notation $\{1\}$ and $\{2\}$ to represent the best and second-best arms, and in general $\{j\}$ represents the j^{th} best arm.

Theorem 8. *Suppose that $\mathbb{J}_t = \mathbb{J}$ for $t \geq 0$ and there exists a single agent who may select any cumulative effort vector process $\boldsymbol{\tau}$ that is $\mathcal{H}_t$-adapted. If $\mathbb{E}[\tau_{\{j\}}(T)] < o(T^a)$ for all $j \neq 1$ and $a > 0$ and $\boldsymbol{\mu} \in \mathbb{R}^J$ with $\mu_{\{1\}} > \mu_{\{2\}}$, then for all $\epsilon > 0$,*

$$\lim_{T \to \infty} \mathbb{P}\left(\tau_{\{2\}}(T) < 2(1-\epsilon)\log T / \left(\mu_{\{1\}} - \mu_{\{2\}}\right)^2\right) = 0.$$

The above theorem states that any policy of the central planner that spends in expectation a subpolynomial amount of time ($o(T^a)$ for all $a > 0$) on each suboptimal arm must also almost surley in the limit spend at least an $2(1-\epsilon)\log T / \left(\mu_{\{1\}} - \mu_{\{2\}}\right)^2$ amount of time on the second-best arm. This implies that any centralized policy with a subpolynomial regret may at best have a logarithmic regret. The proof of Theorem 8 follows the proof of Lai and Robbins [1985]. However, due to the continuous time nature of our model we use a change-of-measure argument that is different from the original Lai and Robbins [1985] proof. Theorem 8 implies if we relaxed the set of available actions of the central planner so that it could replace dropped arms, or even if the central planner could control the actions of the agents, it still could only achieve a regret of order $\Theta(J \log IT)$ over all $\boldsymbol{\mu} \in \boldsymbol{R}^J$. This provides a benchmark regret that we aim for the central planner to achieve when it can only drop arms.

In Section 5 we showed that the regret of the central planner grows linearly if there is a single agent. The discussion of the preceding paragraph now suggests the important question of how many agents are needed for the central planner to mitigate the cost of decentralization and achieve the minimal the regret of $O(JlogT)$? As the time horizon increases, the central planner needs increasingly more agents to achieve the minimal regret. There therefore exists a minimal rate at which the number of agents must grow to achieve the minimal regret. The following negative result states that if the number of agents grows at a rate less than $o(\log T)$, then no policy of the central planner can asymptotically achieve the minimal regret.

Theorem 9. *If the number of agents $I_T = o(\log T)$, then for any policy π of the central planner, $Z_\pi(T) = \omega((I_T T)^a)$ as $T \to \infty$ for all $a \in (0,1)$.*

In order to prove Theorem 9, it suffices to consider the case of two arms. Let π be an arbitrary policy of the central planner and denote by $\boldsymbol{t}_\pi^2 = \left(t_{\pi,1}^2, t_{\pi,2}^2\right)$ the cumulative efforts on arms 1 and 2 under the policy π at the point in time when the central planner discards arm 2. The following lemma states that if the expected cumulative effort on the suboptimal arm $\{2\}$ is less than $o(T^a)$, then arm $\{2\}$ must not be discarded until the cumulative effort on the optimal arm $\{1\}$ is at least $\Omega(\log T)$, that is $t_{\pi,\{1\}}^{\{2\}} = \Omega(\log T)$, $\mathbb{P}$-almost surely.

Lemma 8. *Suppose that $J = 2$ and let π be an arbitrary policy of the central planner. If*

$$\mathbb{E}[\tau_{\pi,\{2\}}^f(T)] = o((IT)^a) \text{ for all } a > 0, \boldsymbol{\mu} \in \mathbb{R}^2,$$

then for all $\epsilon > 0$,

$$\lim_{T\to\infty} \mathbb{P}\left(t_{\pi,\{1\}}^{\{2\}} < 2(1-\epsilon)\log IT / \left(\mu_{\{1\}} - \mu_{\{2\}}\right)^2, t_{\pi,\{2\}}^{\{2\}} < \frac{IT}{2}\right) = 0.$$

We next show that if there are $o(\log T)$ agents, then there is a positive probability that the total cumulative effort on the optimal arm $\{1\}$ is $o(\log T)$ if arm 2 is never discarded which implies that the arm 2 must receive a polynomial expected effort asymptotically.

6.2 The Proposed Policy

We now propose a policy for the central planner that achieves the minimal regret asymptotic of $O(J \log IT)$ for all drift vectors $\boldsymbol{\mu}$ when each arm has a unique drift. We begin by defining a procedure for the central planner to construct confidence bounds for the drift of each arm. Our confidence bounds depend on the amount of effort each agent puts into each arm and therefore may be constructed at any time within the time horizon. Unlike action elimination strategies that make decisions at pre-determined levels of effort for each arm (Even-Dar et al. [2006]), in our policy it is possible for the central planner to use the confidence bounds in order to discard an arm at any point in time and at any level of total effort.

The central planner's ability to make discarding decisions in real time is critical in our setting since the agents cannot be directly controlled and arms may not achieve a pre-determined level of total effort. However, this flexibility comes at the expense of requiring that our confidence bounds guarantee the error introduced due to discarding an arm is bounded at all times. Our error guarantees therefore need to be stronger than the guarantees provided by the confidence bounds in Auer et al. [2002] and Even-Dar et al. [2006]. Note also that an advantage of having multiple agents is that the central planner can use the aggregate information from all agents to make discarding decisions without waiting for agents to settle on an arm, thus speeding up learning.

6.2.1 Confidence Bounds on the Drift

In what follows, we assume the time horizon satisfies $T \geq e\sqrt{\pi/2}$ and the number of arms $J \geq 2$. Suppose now that the central planner has implemented a discarding policy π and that at some point in time t arm $j \in \mathbb{J}$ has a cumulative effort vector $\boldsymbol{s}_j$. We then define the lower and upper confidence bounds on the drift of arm j by

$$C_j^-(\boldsymbol{s}_j) = \frac{R_j^f(\boldsymbol{s}_j)}{\mathbf{1}^T\boldsymbol{s}_j} - \alpha\left(\frac{1}{\mathbf{1}^T\boldsymbol{s}_j} + \frac{1}{\sqrt{\mathbf{1}^T\boldsymbol{s}_j}}\right) \text{ and } \quad C_j^+(\boldsymbol{s}_j) = \frac{R_j^f(\boldsymbol{s}_j)}{\mathbf{1}^T\boldsymbol{s}_j} + \alpha\left(\frac{1}{\mathbf{1}^T\boldsymbol{s}_j} + \frac{1}{\sqrt{\mathbf{1}^T\boldsymbol{s}_j}}\right), \quad (19)$$

where $\alpha = \frac{3}{2}\sqrt{\log\frac{JIT}{\sqrt{2\pi}}}$, which is greater than 3/2 since by assumption $T \geq e\sqrt{\frac{\pi}{2}}$ and $J \geq 2$.

Note that the confidence bounds above are wider than the commonly used confidence bounds in the bandit literature (as in the upper confidence bound or UCB algorithm) because they must be satisfied not just at a fixed level of effort but at all levels of effort. In particular, they have an extra buffer of $\frac{\alpha}{\mathbf{1}^T\boldsymbol{s}_j}$. This buffer prevents the width of the confidence bound from shrinking too fast for low levels of effort (when effort is less than 1), thus preventing the risk of dropping arms without enough evidence. This larger width does not however slow down the learning rate since for large values of effort the extra term $\frac{\alpha}{\sqrt{\mathbf{1}^T\boldsymbol{s}_j}}$ dominates the first term and asymptotically the confidence bounds behave similar to the standard confidence bounds in the bandit literature.

We now characterize the behavior of the confidence bounds above. For each arm $j \in \mathbb{J}$, define the events

$$B_j = \left\{\max_{0<\mathbf{1}^T\boldsymbol{s}_j\leq IT} C_j^-(\boldsymbol{s}_j) > \mu_j\right\} \text{ and } A_j = \left\{\min_{0<\mathbf{1}^T\boldsymbol{s}_j\leq IT} C_j^+(\boldsymbol{s}_j) < \mu_j\right\}.$$

B_j (A_j) are undesirable events that imply the lower (upper) confidence bound for arm j rises above (falls below) its true drift for some amount of effort $0 < s < IT$. Given IT is the maximum effort that any arm can receive across all I agents over the horizon T, if events B_j and A_j do not occur, then arm j's drift always lies within its confidence bounds. The following lemma implies that under any discarding policy π the drift of each arm lies between its confidence bounds at all

times with very high probability.

Lemma 9. *For each arm $j \in \mathbb{J}$,*

$$\mathbb{P}(B_j), \mathbb{P}(A_j) \;<\; \frac{\sqrt{2\pi}\,(\log_4 IT + 2)}{JIT}.$$

We note that large deviation results such as Chernoff bounds are not sufficient to establish Lemma 9 because the event of interest considers entire sample paths rather than the value of confidence bounds at a specific point in time. The bounds in Lemma 9 are instead provided by the escape probability of a Brownian motion from an appropriate non-linear boundary. In order to precisely obtain this probability, one must solve Fokker-Planck equations that do not have a known solution. We instead take a constructive approach of bounding the non-linear boundary by a piecewise linear function. We choose the piecewise linear function carefully to make sure that the bounds are sufficiently tight. We then prove the lemma using the escape probability of a Brownian motion from a piecewise linear boundary, which provides an upper bound on the nonlinear boundary corresponding to the confidence bounds (19).

Lemma 9 together with Bonferroni's inequality implies that

$$P\left(\max_{0<\mathbf{1}^T \boldsymbol{s}_j \leq IT} C_j^-(\boldsymbol{s}_j) < \mu_j < \min_{0<\mathbf{1}^T \boldsymbol{s}_j \leq IT} C_j^+(\boldsymbol{s}_j)\right) > 1 - \frac{2\sqrt{2\pi}\,(\log_4 IT + 2)}{JIT}.$$

We have therefore established that with high probability the true drift of each arm lies between its upper and lower confidence bounds. It turns out that the confidence bounds also converge to the true drift of the arms sufficiently fast. In particular, we have the following result.

Lemma 10. *If events A_j and B_j are false for arm $j \in \mathbb{J}$, then*

$$\mu_j - 2\delta < \min_{\frac{4\alpha^2}{\delta^2} \leq \mathbf{1}^T \boldsymbol{s}_j \leq IT} C_j^-(\boldsymbol{s}_j) \text{ and } \max_{\frac{4\alpha^2}{\delta^2} \leq \mathbf{1}^T \boldsymbol{s}_j \leq IT} C_j^+(\boldsymbol{s}_j) < \mu_j + 2\delta$$

given $2\alpha > \delta$, for all $\delta > 0$.

Lemma 10 states that if the events B_j and A_j do not occur, then the confidence bounds (19) approach the true drift quickly. In particular, for any $\delta > 0$, after time $4\frac{\alpha^2}{\delta^2}$ the confidence bounds are at most 2δ away from the true drift. Lemmas 9 and 10 provide a concentration for the confidence bounds establishing the high probability range for the upper and lower confidence bounds as shown in Figure 3. This is important because if the confidence bounds approach the true drift slowly, then the learning rate is reduced.

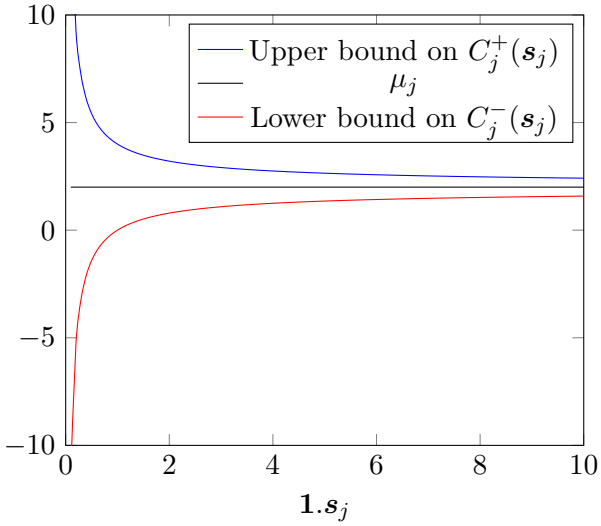

Figure 3: Concentration of the confidence bounds: With probability $1 - \frac{2\sqrt{2\pi}(\log_4 IT+2)}{JIT}$ the upper(lower) confidence bound lies between the blue(red) curve and true drift.

6.2.2 The Proposed Discarding Policy

We now propose a discarding policy for the central planner that asymptotically achieves the minimal regret. We begin by defining the notion of an anchor rate.

Definition 1. Anchor rate: *Given a discarding policy π, the anchor rate at time $0 \leq t \leq T$ is defined to be*

$$l_\pi^*(t) = \max_{j \in \mathbb{J}} \sup_{0<s<t} C_j^- \left(\boldsymbol{\tau}_{\pi,j}(s)\right).$$

The anchor rate is the highest lower confidence bound over all arms up until the current time. It is non-decreasing by definition. Assuming each of the confidence bounds are true, the central planner can safely discard any arm whose upper confidence bounds falls below the anchor rate. We therefore propose the following policy for the central planner.

Definition 2. Central planner's policy π^*: *The proposed policy of the central planner is the unique discarding policy π^* which discards arm $j \in \mathbb{J}$ at time $t > 0$ if*

$$C_j^+ \left(\boldsymbol{\tau}_{\pi^*,j}(t)\right) \leq l_{\pi^*}^*(t).$$

There are two possible sources of error under the policy π^* which could cause its regret to be too large. The first is that the arm with the highest drift could mistkaenly be discarded. The second is that the anchor rate could stay low for too long slowing down the time until the best arm is found. It turns out however that the probability of either of these events is small. We start with the first source of potential error.

If the events B_j and A_j are both false for each $j \in \mathbb{J}$, then under any discarding policy π it follows that $C_j^- \left(\boldsymbol{\tau}_{\pi,j}(s)\right) < \mu_j \leq \mu_{\{1\}}$ for each $j \in \mathbb{J}$ and $s \leq T$, and $C_{\{1\}}^+ \left(\boldsymbol{\tau}_{\pi,\{1\}}(s)\right) > \mu_{\{1\}}$ for $s < T$. Thus, $l_\pi^*(s) < \mu_{\{1\}} < C_{\{1\}}^- \left(\boldsymbol{\tau}_{\pi,j}(s)\right)$ for $s \leq T$. This implies that if B_j and A_j are both false for each $j \in \mathbb{J}$, then under the proposed policy π^* the arm with the highest drift is never discarded. The probability that the arm with the highest drift is discarded under the policy π^* is therefore less than the probability that for one $j \in \mathbb{J}$ either B_j or A_j is true. Using Bonferroni's inequality and Lemma 9, the probability of this event is at most $2\sqrt{2\pi}\left(\log_4 IT + 2\right)/IT$. We state this formally as the following corollary.

Corollary 3. *Under the proposed discarding policy π^*, the probability that the arm with the highest drift is discarded is at most $2\sqrt{2\pi}\left(\log_4 IT + 2\right)/IT$.*

We now provide a bound on the error of the second type. We first state an assumption. This assumption implies that the drifts of all arms can be differentiated.

Assumption 1. *If $l, m \in \mathbb{J}$ such that $l \neq m$, then $\mu_l \neq \mu_m$.*

We refer to the minimum difference in drifts between any pair of arms by 2Δ, that is

$$\Delta = \frac{1}{2} \min_{l,m \in \mathbb{J}, l \neq m} |\mu_l - \mu_m|.$$

Note that by Assumption 1 it follows that $\Delta > 0$. The following theorem states that under Assumption 1, an error of the second type occurs with small probability.

Theorem 10. *Under Assumption 1 and the proposed policy π^*, by time $t = \frac{112\alpha^2}{\Delta^2 I}(J-1)$ the anchor rate satisfies $l^*(t) \geq \mu_{\{1\}} - \Delta$ with probability at least $1 - \frac{2\sqrt{2\pi}(\log_4 IT+2)}{IT} - \log_2 Je^{-\frac{I}{12}}$.*

Theorem 10 provides the time by which the anchor rate rises above the true drift of the second best arm. After this time the effort concentrates on the arms with higher drift. Thus the theorem hints are the learning rate under this policy. Since α^2 grows logarithmically in time T therefore if I grows at least logarithmically in T then the learning occurs in constant time irrespective of the time horizon. The anchor rate rises through successive elimination of poor arms thus concentrating experimentation on better arms. In Figure 4, we demonstrate the lower bound on anchor rate as a function of time. This lower bound progressively improves with time and by time $\frac{112\alpha^2}{\Delta^2 I}(J-1)$ it is above the true drift of the second best arm. The proof of the theorem is delicate. We provide a summary of the proof technique here.

The proof relies on two important properties of the high probability sample paths. The first property is established by Theorems 5 and 6 that says at any given time for each agent the effort spent on the top half (in terms of drift) of any subset of remaining arms stochastically dominates the effort spent on the bottom half of the subset. This property together with the application the Azuma-Hoeffding inequality implies that with high probability, by certain times (corresponding to

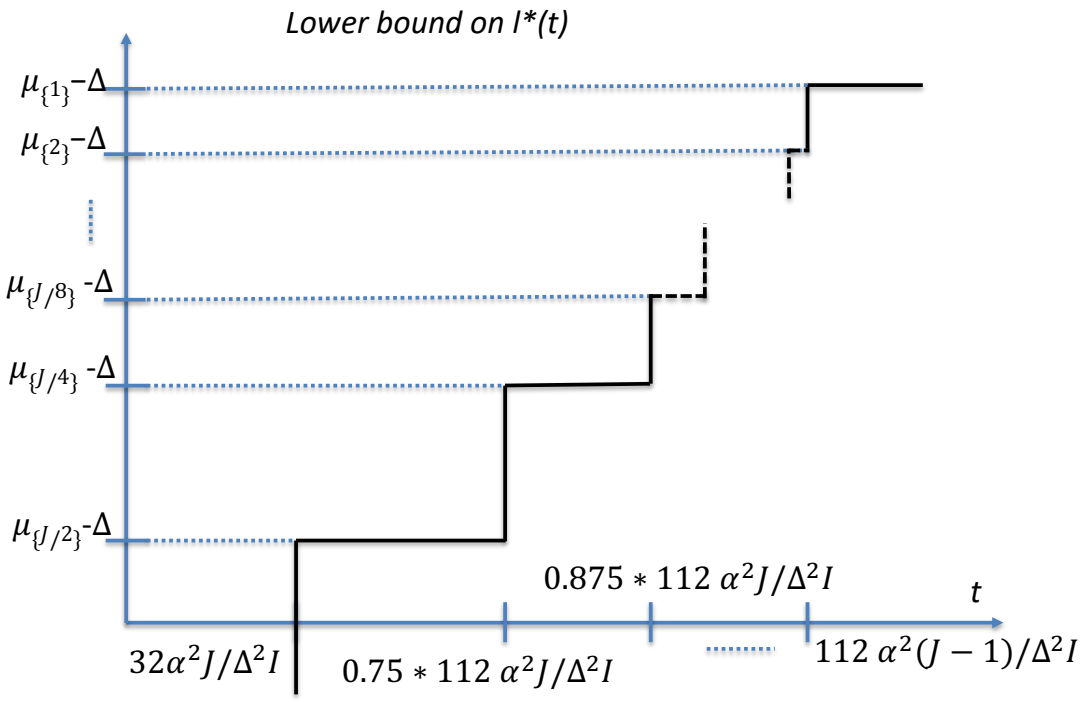

Figure 4: High probability lower bound on the anchor rate.

the jumps in the lower bound on anchor rate), at least one arm in the top half of the remaining arms has received sufficient effort. The second property is established by Lemmas 9 and 10 that says, with high probability, the confidence bounds of arms are well behaved get sufficiently close to their true drift given sufficient effort. Recursive application of these two properties establishes that at time $t = \frac{32\alpha^2 J}{\Delta^2 I}$, the anchor rate is above the true drift of the bottom half of the arms and for $1 \leq n < \log_2 J - 1$, at time $t = \frac{112\alpha^2 J}{\Delta^2 I}(1 - \frac{2^{n-1}}{J})$, the anchor rate is above $\mu_{2^{n-1}} - \Delta$. The recursive argument is delicate. In particular, we do not ignore the effort spent on arms every time the anchor rate rise as it would not lead to a constant learning time. Instead we bound the worst case jump times using a dynamic programming argument to achieve a constant bound on learning time.

6.2.3 Regret under the Proposed Policy

We now present regret bounds on the the proposed policy π^* using the bounds on the probability of errors of the first and second types as stated in Corollary 3 and Theorem 10, respectively. Our first result bounds the expected effort on all suboptimal arms under policy π^*. This bound grows logarithmically in time suggesting it is the rate optimal effort on suboptimal arms.

Theorem 11. *If there are at least $I = 12 \log T$ agents, then under the discarding policy π^*, for any $\boldsymbol{\mu} \in \mathbb{R}^J$ that satisfies Assumption 1, the expected total cumulative effort on suboptimal arms is*

$$\mathbb{E}\left[\sum_{j \neq \{1\}} \tau^f_{\pi^*,j}(T)\right] < \left(\frac{576J}{\Delta^2} + \frac{\sqrt{2\pi}}{\log 2}\right) \log IT + \frac{I \log_2 J}{T^{\frac{I}{12 \log T} - 1}}.$$

As a consequence of the above theorem, using the policy π^* the central planner can achieve a regret of $O(J \log IT)$ uniformly over all $\boldsymbol{\mu} \in \mathbb{R}^J$ that satisfy Assumption 1, assuming the number of agents is at least $12 \log T$. We formally state this in the following corollary.

Corollary 4. *If there are at least $12 \log T$ agents then under policy π^*, for any $\boldsymbol{\mu} \in \mathbb{R}^J$ that satisfies Assumption 1, the regret $\tilde{Z}_{\pi^*}(T, \boldsymbol{\mu})$ is less than*

$$(\mu_{\{1\}} - \mu_{\{J\}}) \left(\left(\frac{576J}{\Delta^2} + \frac{\sqrt{2\pi}}{\log 2}\right) \log IT + \frac{I \log_2 J}{T^{\frac{I}{12 \log T} - 1}}\right).$$

Note that as T grows large, the first term in the above dominates and therefore the expected regret of the central planner under the policy π^* is of order $O(J \log IT)$. Moreover, the total amount of effort spent across all agents up until time T is equal to IT. This may be considered the effective horizon of the central planner. It therefore follows that the regret rate of $\log IT$ is optimal. The above bound on the regret is conservative because it considers the worst possible outcome, where all the effort on suboptimal arms is applied to the arm with the lowest drift. This matches the desired performance bound, suggesting that the policy π^* is optimal up to a constant multiplicative factor.

There is an additional advantage of an increased number of agents for experimentation, it speeds up learning. As Theorem 10 suggests, with very high probability the anchor rate rises above the drift of all suboptimal arms by the time $t = \frac{112\alpha^2}{\Delta^2 I}(J-1)$. This time decreases sharply with I implying that learning speeds up sharply with an increasing number of agents.

6.3 Relaxing the Drift Separation

One natural question to ask is whether Assumption 1 can be relaxed? It turns that when all but the best arm have the same drift, then no feasible policy can achieve the optimal regret bound in the distributed setting with $J \log T$ agents. This implies in our case that as the number of arms increases, the central planner must add proportionally more agents to maintain the same regret order. We state this result in the following theorem.

Theorem 12. *Let π be a discarding policy that for each $\boldsymbol{\mu}$ satisfying Assumption 1 has a regret $\tilde{Z}_\pi(T, \boldsymbol{\mu}) = o(T^a)$ for all $a > 0$ if the number of agents $I_T > 12 \log T$. If $\boldsymbol{\mu}_T$ is such that $\mu_1 = \mu_2 +$*

$\sqrt{1/\log T} = \mu_3 + \sqrt{1/\log T} = \cdots = \mu_J + \sqrt{1/\log T}$ *and the number of agents* $I_T < (1-\beta)J\log T$ *for some* $\beta > 0$*, then the policy* π *has a regret* $\tilde{Z}_\pi(T, \boldsymbol{\mu}_T) = \Omega(T^b)$ *for some* $b \in (0,1)$*.*

The above result implies that any policy having a low regret under Assumption 1 will perform poorly in a particular constructed instance that does not satisfy the rate separation assumption.

7. Conclusion

In this paper, we studied the problem of coordinating learning among a population of independent agents who are not interested in experimentation. We studied this in a multi-armed bandit framework with a twist that the central planner controls the available arms while independent agents make decisions about pulling arms among the ones made available by the central planner. We showed that even if the central planner is never allowed to reinstate removed arms, it can generate enough enough experimentation sufficiently fast, thus mitigating the losses due to decentralization and obtain a regret matching the benchmark under centralized decision making.

We showed that the regret benchmark under centralized decision-making is attainable if the central planner has at least $12\log T$ independent agents and follows a proposed policy that maintains a non-decreasing anchor reward rate and discards arms the moment the lower confidence bound on their reward rates drops below the anchor rate. This forces the agents to experiment and choose among the smaller set of available arms. The ability to distribute experimentation across agents also speeds up learning, thus reducing the burden of experimentation on each agent. For this policy to work well with a small number of agents, we need a slightly stronger notion of differentiation among arms than is needed in the centralized setting.

In our study, we developed new tools and characterizations that are of independent interest for studying other multi-armed bandit settings. In particular, we identified a relationship between the running minimum reward rate estimates of sets of arms and the effort allocated to them. This relationship helps characterize the distribution of efforts of sets of arms using the escape time distribution of a Brownian motion.

Our results have implications for medical experimentation, salesforce management, and platform design, among other applications. Our work also opens up several future directions of investigation. We provide two possible future directions below. First, one can extend our main result to the case when arms are not well differentiated and characterize the tradeoffs between the level of differentiation, number of agents, and the achievable regret. These tradeoffs are important for applications in settings with natural constraints. Second, it would be interesting to study settings with correlated or nonstationary arms or both. Many natural settings do have correlations and nonstationarity and the recent literature on such problems in a classical (centralized) bandit model offers useful benchmarks.

References

Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Learning from reviews: The selection effect and the speed of learning. *Econometrica*, 90(6):2857–2899, 2022.

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

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Yeon-Koo Che and Johannes Hörner. Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics*, 133(2):871–925, 2018.

Bangrui Chen, Peter Frazier, and David Kempe. Incentivizing exploration by heterogeneous users. In *Conference on learning theory*, pages 798–818. PMLR, 2018.

Fangruo Chen. Sales-force incentives and inventory management. *Manufacturing and Service Operations Management*, 2(2):186–202, 2000.

Fangruo Chen. Salesforce Incentives, Market Information, and Production/Inventory Planning. *Management Science*, 51(1):60–75, January 2005.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*, COLT '02, pages 255–270, London, UK, UK, 2002. Springer-Verlag.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning. *Journal Of Machine Learning Research*, 7:1079–1105, 2006.

Peter Frazier, David Kempe, Jon M. Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *EC*, pages 5–22, 2014.

Marina Halac, Navin Kartik, and Qingmin Liu. Optimal contracts for experimentation. *The Review of Economic Studies*, 83(3):1040–1091, 2016.

Eshcar Hillel, Zohar Shay Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed exploration in multi-armed bandits. *CoRR*, abs/1311.0800, 2013.

Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Incentivizing exploration with selective data disclosure. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 647–648, 2020.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101 (6):2590–2615, 2011.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012, 2014.

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

Avi Mandelbaum. Continuous Multi-Armed Bandits and Multiparameter Processes. *Annals of Probability*, 15(4):1527–1556, October 1987.

Avi Mandelbaum and RobertJ. Vanderbei. Brownian bandits. In MarkI. Freidlin, editor, *The Dynkin Festschrift*, volume 34 of *Progress in Probability*, pages 267–285. BirkhÃ¤user Boston, 1994. ISBN 978-1-4612-6691-4.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *J. Mach. Learn. Res.*, 5:623–648, December 2004.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. *Operations Research*, 68(4):1132–1161, 2020.

Yiangos Papanastasiou, Kostas Bimpikis, and Nicos Savva. Crowdsourcing exploration. *Management Science*, 64(4):1727–1746, 2018.

Aleksandrs Slivkins and Eli Upfal. Adapting to a changing environment: the brownian restless bandits. In *COLT*, pages 343–354, 2008.

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Wendler Trickett, Frank Mongiello, and Jordan McLinn. Right to try act of 2017. *One Hundred Fifteenth Congress of the United States of America*, 2017.

P. Whittle. Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):pp. 143–149, 1980.

A. Appendix

A.1 Proofs and Additional Results for Single Agent without a Central Planner: Section 4

A.1.1 Proof of Lemma 1

Proof. We first show that at any finite time the minimum drift estimates of all available products is same almost surely. Assume that there are products $j, l \in \mathbb{J}_t$, and time t such that $L_j(t) < L_l(t)$. This implies that there was a period of time with positive Lebesgue measure for which the inequality was satisfied but the agent spent positive effort on product j. This contradicts the myopic policy and by contradiction the assumption is false. Therefore, at any finite time the minimum drift estimates of all available products is same almost surely.

□

A.1.2 Proof of Theorem 3

Proof. We first note that for all arms $j \neq k^\star$, the cumulative effort $\tau_j(T) = \sigma_j(L(T))$ almost surely. Since, $L(T) \geq \max_{j\in\mathbb{J}} M_j(\infty)$ therefore $\sigma_j(L(T)) \leq \sigma_j(M_j(\infty))$ for all j and therefore the cumulative effort $\tau_j(T) \leq \sigma_j(M_j(\infty))$ for all $j \neq k^\star$. Therefore the expected total cumulative effort on all arms other than $k^\star$ is $\sum_{j\neq k^\star} E[\tau_j(T)] \leq \sum_{j\neq k^\star} E[\sigma_j(M_j(\infty))] = \sum_{j\neq k^\star} c_2(\mu_j)$. Thus this expected effort is upper bounded by a constant independent of time. The cumulative reward of the winning arm $k^\star$ determines the total reward for the myopic agent. The total expected cumulative reward from the winning arm is $\sum_{j=k^\star} T\mu_j P(k^\star = j)$. □

A.1.3 Distribution of the Winning Arm

We now study the expressions for probability that an agent settles on a given arm ultimately.

When all arms have positive drifts: When all arms have positive drift, then the probability is provided by Theorem 4. We provide the proof below.

Proof of Theorem 4

Proof. Assume all arms have positive drifts. The probability

$$\mathbb{P}\left(M_j(\infty) > \max_{m\in\mathbb{J}} M_m(\infty)|\boldsymbol{\mu}\right) = \mathbb{P}\left(M_j(\infty) > \max_{m\in\mathbb{J}} M_m(\infty), M_j(\infty) \leq 0|\boldsymbol{\mu}\right)$$

$$= \mathbb{P}\left(M_j(\infty) - \max_{m\in\mathbb{J}} M_m(\infty) > 0, M_j(\infty) \leq 0|\boldsymbol{\mu}\right).$$

We remind the reader that $M_j(\infty) \leq 0$ always because the drifts estimates at time $t = 0$ is 0. Using the convolution of the distributions to compute the distribution of the sum, the probability

is equal to

$$\int_{-\infty}^{0} e^{2\sum_{m\in\mathbb{J}} x(\mu_m - x)} \frac{d\left(e^{2x(\mu_j - x)}\right)}{dx} dx = \int_{-\infty}^{0} 2\left(\mu_j - 2x\right) e^{2\sum_{m\in\mathbb{J}} x(\mu_m - x)} dx$$

By adding and subtracting $\frac{1}{J-J_0}\sum_{m\in\mathbb{J}}\mu_m$, to $(\mu_j - 2x)$, we obtain the following expression for the probability

$$\int_{-\infty}^{0} \frac{1}{J} 2\left(\sum_{m\in\mathbb{J}} \left(\mu_m - 2x\right)\right) e^{2\sum_{m\in\mathbb{J}} x(\mu_m - x)} dx$$

$$+2\left(\mu_j - \frac{1}{J}\sum_{m\in\mathbb{J}}\mu_m\right)\int_{-\infty}^{0} e^{2\sum_{m\in\mathbb{J}} x(\mu_m - x)} dx$$

The first term in the sum is $\frac{1}{J}$.

This is because

$$\frac{d}{dx} e^{2\sum_{m\in\mathbb{J}} x(\mu_m - x)} = 2\left(\sum_{m\in\mathbb{J}} \left(\mu_m - 2x\right)\right) e^{2\sum_{m\in\mathbb{J}} x(\mu_m - x)}.$$

For the second term, we note that

$$e^{2\sum_{m\in\mathbb{J}} x(\mu_m - x)} = e^{-2J\left(\left(x - \frac{\bar{\mu}}{2}\right)^2 - \frac{\bar{\mu}^2}{4}\right)} = e^{\frac{J\bar{\mu}^2}{2}} e^{-2J\left(x - \frac{\bar{\mu}}{2}\right)^2},$$

where $\bar{\mu} = \frac{1}{J}\sum_{m\in\mathbb{J}}\mu_m$. Therefore, applying the above Gaussian form, we use the Gaussian integral to obtain the following expression for the probability.

$$\frac{1}{J} + \frac{\sqrt{2\pi}\left(\mu_j - \bar{\mu}\right)}{\sqrt{J}} e^{\frac{J\bar{\mu}^2}{2}} \Phi\left(-\sqrt{J}\bar{\mu}\right)$$

□

When some arms have negative drifts: When some arms have negative drift then the expression is quite complex. In the following we, provide the analysis and expression for this probability. For what follows,we will assume that the arms are indexed in the ascending order of their drifts, i.e. $\mu_1 < \mu_2, < \cdots < \mu_J$. We assume that $\mu_{m^\circ} < 0$ and $\mu_{m^\circ+1} > 0$. The probability that an arm j is the winning arm is

$$\mathbb{P}\left(k^\star = j|\boldsymbol{\mu}\right) = \mathbb{P}\left(k^\star = j, L^* \leq \mu_j|\boldsymbol{\mu}\right) + \sum_{l=2}^{j} \mathbb{P}\left(k^\star = j, L^* \in (\mu_l - 1, \mu_l]|\boldsymbol{\mu}\right), \tag{20}$$

where $L^* = \lim_{t\to\infty} L(t)$. We will represent $I_l\left(j\right) := \mathbb{P}\left(k^\star = j, L^* \in (\mu_l - 1, \mu_l]|\boldsymbol{\mu}\right)$. Therefore the

probability can be written as

$$\mathbb{P}\left(k^{\star}=j|\boldsymbol{\mu}\right)=\sum_{l=2}^{j}I_{l}\left(j\right).$$

We will now derive the expressions for each of the terms in the sum. Since L^{*} is at most 0, therefore for any arm j with positive drift and all $m^{\circ}+1<l\leq j$,

$$I_{l}\left(j\right)=0\text{ and }I_{m^{\circ}+1}=\mathbb{P}\left(k^{\star}=j,L^{*}\in(\mu_{m^{\circ}},0]|\boldsymbol{\mu}\right). \tag{21}$$

If j is the winning arm, then the minimum drift estimate of arm j over the whole time horizon must be greater than the minimum drift estimate of all arms over the whole time horizon. Combining this with the fact that the minimum drift estimates of all arms is at most equal to their true drifts almost surely, we obtain the expressions for the probability that any arm j is the winning arm as follows. We use the notation $M_{j}(\infty)=M_{j}(\infty)=\inf_{s>0}\hat{\mu}_{j}(s)$ for simplicity.

$$I_{1}\left(j\right)=\mathbb{P}\left(M_{j}(\infty)>\max_{m\in\mathbb{J}\setminus\{j\}}\{M_{m}(\infty)\},M_{j}(\infty)<\mu_{1}|\boldsymbol{\mu}\right), \tag{22}$$

$$I_{l}\left(j\right)=\mathbb{P}\left(M_{j}(\infty)>\max_{m\in\mathbb{J}\setminus\left(\mathbb{J}_{\mu_{l-1}}\cup\{j\}\right)}\{M_{m}(\infty)\},M_{j}(\infty)\in(\mu_{l}-1,\mu_{l}]|\boldsymbol{\mu}\right),\text{ for all }1<l\leq\min\{m^{\circ},j\} \tag{23}$$

$$I_{m^{\circ}+1}\left(j\right)=\mathbb{P}\left(M_{j}(\infty)>\max_{m\in\mathbb{J}\setminus\left(\mathbb{J}_{\mu_{m^{\circ}}}\cup\{j\}\right)}\{M_{m}(\infty)\},M_{j}(\infty)\in(\mu_{m^{\circ}},0]|\boldsymbol{\mu}\right)\text{ if }j>0, \tag{24}$$

where $\mathbb{J}_{a}=\{m\in\mathbb{J}|\mu_{m}\leq a\}$. The three expressions are very similar expect for the range of L^{*} and the set of arms that are contenders for the winning arm in the range of L^{*}. We first evaluate the following general probability.

$$I\left(j\right)=\mathbb{P}\left(M_{j}(\infty)>\max_{m\in X}\{M_{m}(\infty)\},M_{j}(\infty)\in(a,b]|\boldsymbol{\mu}\right), \tag{25}$$

where $b\leq min\{0,\mu_{j}\}$ and $X=\mathbb{J}\setminus\left(\mathbb{J}_{a}\cup\{j\}\right)$. The three expressions in equations 22, 23 and 24 can be evaluated by replacing a,b and X with appropriate values. The probability $I\left(j\right)$ can be written using convolution as

$$I\left(j\right)=\int_{a}^{b}\mathbb{P}\left(\max_{m\in X}L_{m}\leq x|\boldsymbol{\mu}\right)d\mathbb{P}\left(L_{j}\leq x|\boldsymbol{\mu}\right)$$

Using the expressions for the distribution of minimum drift estimate from Lemma 4, we obtain the following expression for $I\left(j\right)$.

$$I\left(j\right)=\int_{a}^{b}e^{2\sum_{m\in X}x(\mu_{m}-x)}\frac{d\left(e^{2x(\mu_{j}-x)}\right)}{dx}dx$$

Using the derivative of $e^{2x(\mu_j - x)}$ we get the expression

$$I(j) = \int_a^b 2(\mu_j - 2x)\, e^{2\sum_{m\in X\cup\{j\}} x(\mu_m - x)} dx$$

We will refer to the average drift of the arms with the drift greater than a as $\bar{\mu}_a = \frac{1}{|X|+1}\sum_{m\in X\cup\{j\}} \mu_m$. By adding and subtracting $\bar{\mu}_a$ to $(\mu_j - 2x)$we obtain the following expression

$$I(j) = \frac{1}{|X|+1}\int_a^b 2\left(\sum_{m\in X\cup\{j\}} (\mu_m - 2x)\right) e^{2\sum_{m\in X\cup\{j\}} x(\mu_m - x)} dx$$
$$+ 2(\mu_j - \bar{\mu}_a)\int_a^b e^{2\sum_{m\in X\cup\{j\}} x(\mu_m - x)} dx$$

We note that the expression inside the first integral in $I(j)$ is the derivative of the exponential factors in the expression. We note that

$$e^{2\sum_{m\in X\cup\{j\}} x(\mu_m - x)} = e^{-2(|X|+1)\left(\left(x-\frac{\bar{\mu}_a}{2}\right)^2 - \frac{\bar{\mu}_a^2}{4}\right)} = e^{\frac{(|X|+1)\bar{\mu}_a^2}{2}} e^{-2(|X|+1)\left(x-\frac{\bar{\mu}_a}{2}\right)^2}.$$

Using this $I(j)$ is evaluated as

$$I(j) = \frac{1}{|X|+1}\left(e^{2\sum_{m\in X\cup\{j\}} b(\mu_m - b)} - e^{2\sum_{m\in X\cup\{j\}} a(\mu_m - a)}\right) \tag{26}$$
$$+ \frac{\sqrt{2\pi}(\mu_j - \bar{\mu}_a)}{\sqrt{|X|+1}} e^{\frac{(|X|+1)\bar{\mu}_a^2}{2}}\left(\Phi\left(2\sqrt{|X|+1}\left(b - \frac{\bar{\mu}_a}{2}\right)\right) - \Phi\left(2\sqrt{|X|+1}\left(a - \frac{\bar{\mu}_a}{2}\right)\right)\right) \tag{27}$$

Replacing a, b and X with appropriate values, we get the three joint probability expressions. For $I_1(j)$, we replace a with $-\infty$, b with μ_1 and X with $\mathbb{J}\setminus\{j\}$. We get the following expression for $I_1(j)$.

$$I_1(j) = \frac{1}{J} e^{2\sum_{m\in\mathbb{J}} \mu_1(\mu_m - \mu_1)} + \frac{\sqrt{2\pi}(\mu_j - \bar{\mu})}{\sqrt{J}} e^{\frac{J\bar{\mu}^2}{2}} \Phi\left(2\sqrt{J}\left(\mu_1 - \frac{\bar{\mu}}{2}\right)\right), \tag{28}$$

where $\bar{\mu} = \frac{1}{J}\sum_{m\in\mathbb{J}} \mu_m$. Replacing a, b and X with appropriate values, we get the three joint probability expressions. For all $1 < l \leq \min\{m^\circ, j\}$, for the expression for $I_l(j)$, we replace a with μ_{l-1}, b with μ_l and X with $\mathbb{J}\setminus\left(\mathbb{J}_{\mu_{l-1}}\cup\{j\}\right)$. We get the following expression for $I_l(j)$.

$$I_l(j) = \frac{1}{J-l+1}\left(e^{2\sum_{m\in\mathbb{J}\setminus\mathbb{J}_{\mu_{l-1}}} \mu_l(\mu_m - \mu_l)} - e^{2\sum_{m\in\mathbb{J}\setminus\mathbb{J}_{\mu_{l-1}}} \mu_{l-1}(\mu_m - \mu_{l-1})}\right) \tag{29}$$

$$+ \frac{\sqrt{2\pi}(\mu_j - \bar{\mu}_l)}{\sqrt{J-l+1}} e^{\frac{(J-l+1)\bar{\mu}_l^2}{2}}\left(\Phi\left(2\sqrt{J-l+1}\left(\mu_l - \frac{\bar{\mu}_l}{2}\right)\right) - \Phi\left(2\sqrt{J-l+1}\left(\mu_{l-1} - \frac{\bar{\mu}_l}{2}\right)\right)\right),$$

where $\bar{\mu}_l = \frac{1}{J-l}\sum_{m\in\mathbb{J}\setminus\mathbb{J}_{\mu_l}} \mu_m$. Replacing a, b and X with appropriate values, we get the three joint

probability expressions. For the expression for $I_{m^\circ+1}(j)$, we replace a with μ_{m°, b with 0 and X with $\mathbb{J} \setminus \left(\mathbb{J}_{\mu_{m^\circ}} \cup \{j\}\right)$. We get the following expression for $I_{m^\circ+1}(j)$.

$$I_{m^\circ+1}(j) = \frac{1}{J-m^\circ}\left(1 - e^{2\sum_{m\in\mathbb{J}\setminus\mathbb{J}_{\mu_{m^\circ}}} \mu_{m^\circ}(\mu_m - \mu_{m^\circ})}\right) \tag{30}$$

$$+\frac{\sqrt{2\pi}\left(\mu_j - \bar{\mu}_{m^\circ}\right)}{\sqrt{J-m^\circ}} e^{\frac{(J-m^\circ)\bar{\mu}_{m^\circ}^2}{2}} \left(\Phi\left(-2\sqrt{J-m^\circ}\left(\frac{\bar{\mu}_{m^\circ}}{2}\right)\right) - \Phi\left(2\sqrt{J-m^\circ}\left(\mu_{m^\circ} - \frac{\bar{\mu}_{m^\circ}}{2}\right)\right)\right),$$

where $\bar{\mu}_{m^\circ} = \frac{1}{J-m^\circ}\sum_{m\in\mathbb{J}\setminus\mathbb{J}_{\mu_{m^\circ}}} \mu_m$.

A.2 Proofs and Additional Results for the Case with the Central Planner and a Single Agent: Section 5

A.2.1 Proof of Lemma 7

Proof. We will prove the first statement by contradiction. Assume $L_{\pi_K}(t^\circ) > L_\pi(t)$. This implies that $\sum_{j\in K}\sigma_j\left(L_\pi(t)\right) > t^\circ$ according to Lemma 5. Since $L_{\pi,j}(t) = L_\pi(t)$ for all $j \in K$ almost surely, by Lemma 1, therefore $\tau_{\pi,j}(t) \geq \sigma_j\left(L_\pi(t)\right)$ for all $j \in K$ almost surely. This implies that $\sum_{j\in K}\tau_{\pi,j}(t) \geq \sum_{j\in K}\sigma_j\left(L_\pi(t)\right) > t^\circ$ almost surely which contradicts Equation 16.

Now assume $L_{\pi_K}(t^\circ) < L_\pi(t)$. Therefore, there exists $\epsilon > 0$ such that $L_{\pi_K}(t^\circ) = L_\pi(t) - \epsilon$. This implies that $\sum_{j\in K}\sigma_j\left(L_\pi(t) - \epsilon\right) \leq t^\circ$ according to Lemma 5. Since $L_{\pi,j}(t) = L_\pi(t) > L_\pi(t) - \epsilon$ for all $j \in K$ almost surely, by Lemma 1, therefore $\tau_{\pi,j}(t) < \sigma_j\left(L_\pi(t) - \epsilon\right)$ for all $j \in K$ almost surely. This implies that $\sum_{j\in K}\tau_{\pi,j}(t) < \sum_{j\in K}\sigma_j\left(L_{\pi_K}(t^\circ)\right) \leq t^\circ$ almost surely again contradicting Equation 16. This completes the proof of the first statement.

By Lemma 2 almost surely $k_\pi(t)$ is not empty. Without loss of generality assume that $j \in k_\pi(t)$. Then by Lemma 2 almost surely j is the only product in $k_\pi(t)$. Also by Lemma 1 $L_{\pi,l}(t) = L_\pi(t)$ for all $l \in K$, by Corollary 1 $\tau_{\pi,l}(t) = \sigma_l\left(L_\pi(t)\right) = \sigma_l\left(L_{\pi_K}(t^\circ)\right)$, for all $l \in K \setminus \{j\}$, and by Theorem 1, $\underline{\mu}_l(s) < L_{\pi,j}(t) = L_{\pi_K}(t^\circ)$ for all $s > \tau_{\pi,l}(t)$ and for all $l \in K \setminus \{j\}$ almost surely. This implies that for all products $l \in K \setminus \{j\}$, $\tau_{\pi_K,l}(t^\circ) = \sigma_l\left(L_{\pi_K}(t^\circ)\right) = \sigma_l\left(L_\pi(t)\right) = \tau_{\pi,l}(t)$ almost surely. This implies that $\tau_{\pi_K,j}(t^\circ) = t^\circ - \sum_{l\in K\setminus\{j\}}\tau_{\pi_K,l}(t^\circ) = t^\circ - \sum_{l\in K\setminus\{j\}}\tau_{\pi,l}(t) = \tau_{\pi,j}(t)$ and $j \in k_{\pi_K}(t^\circ)$ almost surely. This completes the proof of the second and third statement.

□

A.2.2 Proof of Theorem 5

Proof. Assume $k \subseteq \mathbb{J}_t$. We point that by lemma 7 given $\sum_{j\in K}\tau_{\pi,j}(t) = t^\circ$, $\tau_{\pi,j_1}(t^\circ) = \tau_{\pi_K,j_1}(t^\circ)$ and $\tau_{\pi,j_2}(t^\circ) = \tau_{\pi_K,j_2}(t^\circ)$ almost surely and by lemma 12 $\tau_{\pi_K,j_1}(t^\circ)$ first-order stochastically dom-

inates $\tau_{\pi_K,j_2}(t^\circ)$. Therefore, for all $0 < s < t^\circ$,

$$\mathbb{P}\left(\tau_{\pi,j_1}(t) > s | \sum_{j\in K}\tau_{\pi,j}(t) = t^\circ\right) = \mathbb{P}\left(\tau_{\pi_K,j_1}(t^\circ) > s\right)$$

$$> \mathbb{P}\left(\tau_{\pi_K,j_2}(t^\circ) > s\right) = \mathbb{P}\left(\tau_{\pi,j_2}(t) > s | \sum_{j\in K}\tau_{\pi,j}(t) = t^\circ\right).$$

□

A.2.3 Proof of Theorem 6

Proof. We will show that for any $j \in K \cap \mathbb{J}_t$ and $x \geq 0$, $\tau_{\pi_K,j}(t^\circ) > x$ implies that given $\sum_{l\in K}\tau_{\pi,l}(t) = t^\circ$, $\tau_{\pi,j}(t) > x$ almost surely. This will imply that

$$\mathbb{P}\left(\tau_{\pi_K,j}(t^\circ) > x\right) \leq \mathbb{P}\left(\tau_{\pi,j}(t) > x \,|\, \sum_{l\in K}\tau_{\pi,l}(t) = t^\circ\right)$$

proving the result. To prove, we assume that $\tau_{\pi_K,j}(t^\circ) > x$. By lemma 11, this implies that $L_{\pi_{\{j\}}}(x) > L_{\pi_{K\setminus\{j\}}}(t^\circ - x)$ almost surely. Since $L_{\pi_{\{j\}}}(x)$ and $L_{\pi_{K\setminus\{j\}}}(t^\circ - x)$ have continuous distributions and they are independent therefore $L_{\pi_{\{j\}}}(x) \geq L_{\pi_{K\setminus\{j\}}}(t^\circ - x)$ almost surely. Therefore according to Lemma 5 $\sum_{l\in K\setminus\{j\}}\sigma_l\left(L_{\pi_{\{j\}}}(x)\right) \leq t^\circ - x$ almost surely. Since $\sigma_l(y)$ has a continuous distribution for all l, y, therefore $\sum_{l\in K\setminus\{j\}}\sigma_l\left(L_{\pi_{\{j\}}}(x)\right) < t^\circ - x$ almost surely.

From Corollary 2, for all $m \notin K \setminus \mathbb{J}_t$, $L_{\pi,m}(t) \leq L_{\pi,j}(t)$ almost surely. Since, $\tau_{\pi_K,j}(t^\circ) > x$ therefore $L_{\pi,j}(t) = L_{\pi_{\{j\}}}(\tau_{\pi_K,j}(t^\circ)) \leq L_{\pi_{\{j\}}}(x)$ almost surely. Therefore, $L_{\pi,m}(t) \geq L_{\pi_{\{j\}}}(x)$ almost surely. Therefore, $\tau_{\pi,m}(t) \leq \sigma_m\left(L_{\pi_{\{j\}}}(x)\right)$ almost surely. This implies that almost surely

$$\begin{aligned}
\sum_{l\in K\cap\mathbb{J}_t\setminus\{j\}}\sigma_l\left(L_{\pi_{\{j\}}}(x)\right) &= \sum_{l\in K\setminus\{j\}}\sigma_l\left(L_{\pi_{\{j\}}}(x)\right) - \sum_{l\in K\setminus(\mathbb{J}_t\cup\{j\})}\sigma_l\left(L_{\pi_{\{j\}}}(x)\right)\\
&< t^\circ - x - \sum_{m\in K\setminus(\mathbb{J}_t\cup\{j\})}\sigma_m\left(L_{\pi_{\{j\}}}(x)\right)\\
&\leq t^\circ - x - \sum_{m\in K\setminus(\mathbb{J}_t\cup\{j\})}\tau_{\pi_m}(t).
\end{aligned}$$

Therefore by lemma 5,

$$L_{\pi_{\{j\}}}(x) > L_{\pi_{K\cap\mathbb{J}_t\setminus\{j\}}}\left(t^\circ - x - \sum_{m\in K\setminus(\mathbb{J}_t\cup\{j\})}\tau_m(t)\right)$$

almost surely. Therefore by lemma 11

$$\tau_{\pi_{K\cap\mathbb{J}_t},j}\left(t^{\circ}-\sum_{m\in K\setminus(\mathbb{J}_t\cup\{j\})}\tau_{\pi,m}\left(t\right)\right)>x$$

almost surely. Therefore by lemma 7, $\tau_{\pi,j}\left(t\right)>x$, given $\sum_{l\in K\cap\mathbb{J}_t}\tau_{\pi,l}\left(t\right)=t^{\circ}-\sum_{m\in K\setminus(\mathbb{J}_t\cup\{j\})}\tau_{\pi,m}\left(t\right)$ almost surely. This implies that $\tau_{\pi,j}\left(t\right)>x$, given $\sum_{l\in K}\tau_{\pi,l}\left(t\right)=t^{\circ}$ almost surely. This completes the proof. □

A.2.4 Effort Allocations for a Restricted Set of Arms

We characterize the cumulative effort distributions for different arms under π_K policies for any $K\subseteq\mathbb{J}$. These results are useful for proving Theorems 5 and 6.

Lemma 11. *For any arm $j\in K\subseteq\mathbb{J}$ and $x>0$, $\tau_{\pi_K,j}\left(t\right)>x$ if and only if $L_{\pi_{\{j\}}}\left(x\right)>L_{\pi_{K\setminus\{j\}}}\left(t-x\right)$ almost surely.*

Proof. To prove the lemma, we check two cases:

(i) If $j\in k_{\pi_K}\left(t\right)$ and $\tau_{\pi_K,j}\left(t\right)>x$ then by Theorem 1 $L_{\pi_{\{l\}}}\left(y\right)<L_{\pi_K}\left(t\right)$ for all $l\in K\setminus\{j\}$ and $y>\tau_{\pi_K,l}\left(t\right)$ and $L_{\pi_{\{j\}}}\left(x\right)\geq L_{\pi_K}\left(t\right)$ almost surely. This implies that $L_{\pi_{K\setminus\{j\}}}\left(t-x\right)<L_{\pi_K}\left(t\right)\leq L_{\pi_{\{j\}}}\left(x\right)$ almost surely. On the other hand if $j\in k_{\pi_K}\left(t\right)$ and $\tau_{\pi_K,j}\left(t\right)\leq x$ then by Theorem 1 $L_{\pi_{\{l\}}}\left(y\right)\geq L_{\pi_K}\left(t\right)$ for all $l\in K\setminus\{j\}$ and $y\leq\tau_{\pi_K,l}\left(t\right)$ and $L_{\pi_{\{j\}}}\left(x\right)\leq L_{\pi_K}\left(t\right)$ almost surely. This implies that $L_{\pi_{K\setminus\{j\}}}\left(t-x\right)\geq L_{\pi_K}\left(t\right)\geq L_{\pi_{\{j\}}}\left(x\right)$ almost surely.

(ii) If $j\notin k_{\pi_K}\left(t\right)$ and $\tau_{\pi_K,j}\left(t\right)>x$ then from theorem 1 $L_{\pi_{\{j\}}}\left(x\right)>L_{\pi_K}\left(t\right)$ and $L_{\pi_{K\setminus\{j\}}}\left(t-x\right)\leq L_{\pi_K}\left(t\right)$ almost surely. This implies that $L_{\pi_{K\setminus\{j\}}}\left(t-x\right)\leq L_{\pi_K}\left(t\right)<L_{\pi_{\{j\}}}\left(x\right)$ almost surely. On the other hand, if $j\notin k_{\pi_K}\left(t\right)$ and $\tau_{\pi_K,j}\left(t\right)\leq x$ then by Theorem 1 $L_{\pi_{\{j\}}}\left(x\right)\leq L_{\pi_K}\left(t\right)$ and $L_{\pi_{K\setminus\{j\}}}\left(t-x\right)\geq L_{\pi_K}\left(t\right)$ almost surely. This implies that $L_{\pi_{K\setminus\{j\}}}\left(t-x\right)\geq L_{\pi_K}\left(t\right)\geq L_{\pi_{\{j\}}}\left(x\right)$ almost surely. □

We now apply this Lemma to compare the distribution of cumulative efforts for two different arms.

Lemma 12. *Given two arms $j_1,j_2\in K\subseteq\mathbb{J}$, with $\mu_{j_1}>\mu_{j_2}$, $\tau_{\pi_K,j_1}\left(t\right)$ first-order stochastically dominates $\tau_{\pi_K,j_2}\left(t\right)$.*

Proof. By Lemma 11, for any $y>0$,

$$\mathbb{P}\left(\tau_{\pi_K,j_1}\left(t\right)>y\right)=\mathbb{P}\left(L_{\pi_{\{j_1\}}}\left(y\right)-L_{\pi_{K\setminus\{j_1\}}}\left(t-y\right)>0\right)$$

and

$$\mathbb{P}\left(\tau_{\pi_K,j_2}\left(t\right)>y\right)=\mathbb{P}\left(L_{\pi_{\{j_2\}}}\left(y\right)-L_{\pi_{K\setminus\{j_2\}}}\left(t-y\right)>0\right).$$

We note that by Lemma 4, $L_{\pi_{A(\{j_1\}}}(y)$ first-order stochastically dominates $L_{\pi_{\{j_2\}}}(y)$ and by lemma 6 $L_{\pi_{K\setminus\{j_2\}}}(t-x)$ first-order stochastically dominates $L_{\pi_{K\setminus\{j_1\}}}(t-x)$. Therefore,

$$\mathbb{P}\left(L_{\pi_{\{j_1\}}}(y) - L_{\pi_{K\setminus\{j_1\}}}(t-y) > 0\right) > \mathbb{P}\left(L_{\pi_{\{j_2\}}}(y) - L_{\pi_{K\setminus\{j_2\}}}(t-y) > 0\right).$$

This implies that for any $x > 0$, $\mathbb{P}\left(\tau_{\pi_K,j_1}(t) > x\right) > \mathbb{P}\left(\tau_{\pi_K,j_2,}(t) > x\right)$ completing the proof. □

A.3 Proofs for Section 6

A.3.1 Proof of Theorem 8

Proof. Consider two expected profit rate vectors $\boldsymbol{\mu}^{\circ} = (\mu_1^{\circ}, \mu_2)$ and $\boldsymbol{\mu}^{*} = (\mu_1^{*}, \mu_2)$ where $\mu_1^{\circ} < \mu_2 < \mu_1^{*}$ with $(\mu_2 - \mu_1^{\circ}) = \delta(\mu_1^{*} - \mu_2)$ for some $\delta \in (0,1)$. From the assumption, the

$$\boldsymbol{E}_{\boldsymbol{\mu}^{*}}[\tau_{\pi,2}(T)] < o(T^a)$$

for all $0 < a < \delta$. Consider the event

$$X_T \equiv \{\tau_{\pi,1}(T) < 2(1-\delta)\log T/(\mu_1^{*} - \mu_1^{\circ})^2\}.$$

$$\mathbb{P}_{\boldsymbol{\mu}^{*}}(X_T)\left(T - 2(1-\delta)\log T/(\mu_1^{*} - \mu_1^{\circ})^2\right) < \boldsymbol{E}_{\boldsymbol{\mu}^{*}}[\tau_{\pi,2}(T)].$$

Therefore,

$$\mathbb{P}_{\boldsymbol{\mu}^{*}}(X_T) = o\left(T^{a-1}\right).$$

Using the change of measure, the Radon-Nikodyn derivative

$$\ell(\tau_{\pi,1}(T)) = \frac{\partial \mathbb{P}_{\boldsymbol{\mu}^{\circ}}(\tau_{\pi,1}(T))}{\partial \mathbb{P}_{\boldsymbol{\mu}^{*}}(\tau_{\pi,1}(T))} = e^{\left(\mu_1^{*} - \mu_1^{\circ}\right)B_1(\tau_{\pi,1}(T)) + \frac{(\mu_1^{*} - \mu_1^{\circ})^2 \tau_{\pi,1}(T)}{2}},$$

where $B_1(\tau_1(T))$ is the value of the Brownian motion corresponding to the reward process off product 1 for the total cumulative effort of $\tau_{\pi 1}(T)$. Therefore, the probability

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\mu}^{\circ}}\left(X_T, \ell(\tau_{\pi,1}(T)) < T^{1-a}\right) = \int_{X_T \cap \ell(\tau_{\pi,1}(T)) < T^{1-a}} \ell(\tau_{\pi,1}(T)) \partial \mathbb{P}_{\boldsymbol{\mu}^{*}}(\tau_{\pi,1}(T)) \\
&< \int_{X_T \cap \ell(\tau_{\pi,1}(T)) < T^{1-a}} T^{1-a} \partial \mathbb{P}_{\boldsymbol{\mu}^{*}}(\tau_{\pi,1}(T)) = T^{1-a}\mathbb{P}_{\boldsymbol{\mu}^{*}}\left(X_T, \ell(\tau_{\pi,1}(T)) < T^{1-a}\right) \\
&< T^{1-a}\mathbb{P}_{\boldsymbol{\mu}^{*}}(X_T) = T^{1-a} o\left(T^{a-1}\right).
\end{aligned}$$

Therefore,

$$\lim_{T\to\infty} \mathbb{P}_{\boldsymbol{\mu}^{\circ}}\left(X_T, \ell(\tau_{\pi,1}(T)) < T^{1-a}\right) = 0.$$

We note that

$$\lim_{t'\to\infty}\frac{\log \ell(\tau_{\pi,1}(T))}{t'}=\lim_{t'\to\infty}\left(\mu_1^*-\mu_1^\circ\right)\frac{B_1\left(\tau_{\pi,1}(T)\right)}{t'}+\frac{\left(\mu_1^*-\mu_1^\circ\right)^2\tau_{\pi,1}(T)}{2t'}=\frac{\left(\mu_1^*-\mu_1^\circ\right)^2\tau_{\pi,1}(T)}{2t'},$$

for all $t'\geq\tau_{\pi,1}(T)$ $a.s.$ Therefore replacing t' by $2\left(1-\delta\right)\log T/\left(\mu_1^*-\mu_1^\circ\right)^2$ we obtain $\lim_{T\to\infty}\ell(\tau_{\pi,1}(T))\leq T^{1-\delta}<T^{1-a}$ for all $\tau_{\pi,1}(T)\leq\left(2\left(1-\delta\right)\log T/\left(\mu_1^*-\mu_1^\circ\right)^2\right)$ almost surely. This implies that

$$\lim_{T\to\infty}\mathbb{P}_{\boldsymbol{\mu}^\circ}\left(X_T\right)=0.$$

Therefore

$$\lim_{T\to\infty}\mathbb{P}\left(\tau_{\pi,\{2\}}(T)<\frac{2\left(1-\delta\right)\log T}{\left(1+\delta\right)^2\left(\mu_{\{1\}}-\mu_{\{2\}}\right)^2}\right)=0.$$

Since $\delta>a$ and a is arbitrarily chosen between $0,1$, the claim follows. □

A.3.2 Proofs of Theorem 9 and Lemma 8

We first give the proof of Lemma 8.

Proof of Lemma 8. Consider two expected profit rate vectors $\boldsymbol{\mu}^\circ=(\mu_1^\circ,\mu_2)$ and $\boldsymbol{\mu}^*=(\mu_1^*,\mu_2)$ where $\mu_1^\circ<\mu_2<\mu_1^*$ with $\left(\mu_2-\mu_1^\circ\right)=\delta\left(\mu_1^*-\mu_2\right)$ for some $\delta\in(0,1)$. From the assumption, under the policy π, the

$$\boldsymbol{E}_{\boldsymbol{\mu}^\circ}[\tau_{\pi,1}^f\left(T\right)]<o\left((IT)^a\right)$$

for all $0<a<\delta$. Consider the event

$$X_T\equiv\{t_{\pi,1}^2<2\left(1-\delta\right)\log IT/\left(\mu_1^*-\mu_1^\circ\right)^2,\text{ and }t_{\pi,2}^2<\frac{IT}{2}\}.$$

$$\mathbb{P}_{\boldsymbol{\mu}^\circ}\left(X_T\right)\left((IT)/2\right)<\boldsymbol{E}_{\boldsymbol{\mu}^\circ}[\tau_{\pi,1}^f\left(T\right)].$$

Therefore,

$$\mathbb{P}_{\boldsymbol{\mu}^\circ}\left(X_T\right)=o\left((IT)^{a-1}\right).$$

Using the change of measure, the Radon-Nikodyn derivative

$$\ell(\boldsymbol{t}_\pi^2)=\frac{\partial\mathbb{P}_{\boldsymbol{\mu}^*}\left(\boldsymbol{t}_\pi^2\right)}{\partial\mathbb{P}_{\boldsymbol{\mu}^\circ}\left(\boldsymbol{t}_\pi^2\right)}=e^{-\left(\mu_1^*-\mu_1^\circ\right)B_1\left(t_{\pi,1}^2\right)+\frac{\left(\mu_1^*-\mu_1^\circ\right)^2t_{\pi,1}^2}{2}},$$

where $B_1\left(t_{\pi,1}^2\right)$ is the value of the Brownian motion corresponding to the reward process off product

1 for the total cumulative effort of $t^2_{\pi,1}$. Therefore, the probability

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\mu}^*}\left(X_T, \ell(\boldsymbol{t}^2_\pi) < (IT)^{1-a}\right) = \int_{X_T \cap \ell(\boldsymbol{t}^2_\pi)<(IT)^{1-a}} \ell(\boldsymbol{t}^2_\pi) \partial \mathbb{P}_{\boldsymbol{\mu}^\circ}\left(\boldsymbol{t}^2_\pi\right) \\
&< \int_{X_T \cap \ell(\boldsymbol{t}^2_\pi)<(IT)^{1-a}} (IT)^{1-a} \partial \mathbb{P}_{\boldsymbol{\mu}^\circ}\left(\boldsymbol{t}^2_\pi\right) = (IT)^{1-a} \mathbb{P}_{\boldsymbol{\mu}^\circ}\left(X_T, \ell(\boldsymbol{t}^2_\pi) < (IT)^{1-a}\right) \\
&< (IT)^{1-a} \mathbb{P}_{\boldsymbol{\mu}^\circ}\left(X_T\right) = (IT)^{1-a} o\left((IT)^{a-1}\right).
\end{aligned}$$

Therefore,

$$\lim_{T\to\infty} \mathbb{P}_{\boldsymbol{\mu}^*}\left(X_T, \ell(\boldsymbol{t}^2_\pi) < (IT)^{1-a}\right) = 0.$$

We note that

$$\lim_{t'\to\infty} \frac{\log \ell(\boldsymbol{t}^2_\pi)}{t'} = \lim_{t'\to\infty} -\left(\mu^*_1 - \mu^\circ_1\right) \frac{B_1\left(t^2_{\pi,1}\right)}{t'} + \frac{\left(\mu^*_1 - \mu^\circ_1\right)^2 t^2_{\pi,1}}{2t'} = \frac{\left(\mu^*_1 - \mu^\circ_1\right)^2 t^2_{\pi,1}}{2t'} \text{ for all } t' \geq t^2_{\pi,1} \ a.s.$$

Therefore replacing t' by $2\left(1-\delta\right)\log IT/\left(\mu^*_1 - \mu^\circ_1\right)^2$ we obtain
$\lim_{T\to\infty} \ell(\boldsymbol{t}^2_\pi) \leq (IT)^{1-\delta} < (IT)^{1-a}$ for all $\boldsymbol{t}^2_\pi \leq (2\left(1-\delta\right)\log IT/\left(\mu^*_1 - \mu^\circ_1\right)^2, \frac{IT}{2})$ a.s. This implies that

$$\lim_{T\to\infty} \mathbb{P}_{\boldsymbol{\mu}^*}\left(X_T\right) = 0.$$

Therefore

$$\lim_{T\to\infty} \mathbb{P}\left(t^{\{2\}}_{\pi,\{1\}} < \frac{2\left(1-\delta\right)\log IT}{\left(1+\delta\right)^2\left(\mu_{\{1\}} - \mu_{\{2\}}\right)^2}, t^{\{2\}}_{\pi,\{2\}} < \frac{IT}{2}\right) = 0.$$

Since δ is arbitrarily chosen between $0, 1$, the claim follows. □

We now give the proof of Theorem 9

Proof of Theorem 9. We again consider the case of two products with $\boldsymbol{\mu} = (\mu_1, \mu_2)$, $\mu_1 > \mu_2$ and a sequence of settings indexed by time T with I_T agents. We first point out that $Z_\pi(T) = O(\log(I_T T))$ then $Z_\pi(T) = o(T^a)$ for all $a > 0$. Therefore by lemma 8 any policy π with $Z_\pi(T) = O(\log(I_T T))$ must have

$$\lim_{T\to\infty} \mathbb{P}\left(t^2_{\pi,1} < 2\left(1-\epsilon\right)\log\left(I_T T\right)/\left(\mu_1 - \mu_2\right)^2, t^2_{\pi,2} < \frac{I_T T}{2}\right) = 0.$$

Therefore, we can restrict the candidate policies to the set policies that satisfy this condition. Asymptotically, such policies do not discard product 2 before $T/2$ unless

$\tau^f_{\pi,1}(T/2) > 2\left(1-\epsilon\right)\log\left(I_T T\right)/\left(\mu_1 - \mu_2\right)^2$ almost surely. Such policies also discard product 1 before time $T/2$ with probability at most $o((I_T T)^{a-1})$ for all $0 < a < 1$ otherwise the expected

regret for this $\boldsymbol{\mu}$ would be $\Omega((I_T T)^a)$. This implies that for $0 < a < 1$ by setting $\epsilon = 1/2$

$$\begin{aligned}
&\lim_{T\to\infty} \mathbb{P}\left(\tau^f_{\pi,2}(T) > \Omega((I_T T)^a)\right) \\
&> \mathbb{P}\left(\tau^f_{\pi,2}(T/2) > \Omega((I_T T)^a), \mathbb{J}_{T/2} = \{1,2\}\right) + \lim_{T\to\infty} \mathbb{P}\left(1 \notin \mathbb{J}_{T/2}\right) \\
&> \lim_{T\to\infty} \mathbb{P}\left(\tau^f_{\pi,1}(T/2) < 2\left(1-\epsilon\right)\log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2, \mathbb{J}_{T/2} = \{1,2\}\right) + \lim_{T\to\infty} \mathbb{P}\left(1 \notin \mathbb{J}_{T/2}\right) \\
&= \lim_{T\to\infty} \mathbb{P}\left(\tau^f_{1,A(\{1,2\})}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2, 1 \in \mathbb{J}_{T/2}\right) + \lim_{T\to\infty} \mathbb{P}\left(1 \notin \mathbb{J}_{T/2}\right)
\end{aligned}$$

The last equality applies because $\tau^f_{\pi,1}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2$ almost surely implies that $2 \in \mathbb{J}_{T/2}$ and

$$\tau^f_{\pi,1}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2 \Leftrightarrow \tau^f_{\pi_{\{1,2\}},1}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2$$

almost surely following lemma 7. Therefore the above probability is greater than

$$\begin{aligned}
&\lim_{T\to\infty} \mathbb{P}\left(\tau^f_{\pi_{\{1,2\}},1}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2\right) - \lim_{T\to\infty} \mathbb{P}\left(1 \notin \mathbb{J}_{T/2}\right) + \lim_{T\to\infty} \mathbb{P}\left(1 \notin \mathbb{J}_{T/2}\right) \\
&= \lim_{T\to\infty} \mathbb{P}\left(\tau^f_{\pi_{\{1,2\}},1}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2\right).
\end{aligned}$$

From Lemma 11 for any agent i and $x < T/2$,

$$\begin{aligned}
&\mathbb{P}\left(\tau^i_{\pi_{\{1,2\}},1}(T/2) < x\right) \\
&= \mathbb{P}\left(L^i_{\pi,1}\left(x\right) < L^i_{\pi,2}\left(T/2-x\right)\right) \\
&> \mathbb{P}\left(L^i_{\pi,1}\left(x\right) < \left(\mu_2-y\right), L^i_{\pi,2}\left(T/2-x\right) > \left(\mu_2-y\right)\right) \text{ for all } y > \mu_2 \\
&= \mathbb{P}\left(L^i_{\pi,1}\left(x\right) < \left(\mu_2-y\right)\right)\mathbb{P}\left(L^i_{\pi,2}\left(T/2-x\right) > \left(\mu_2-y\right)\right).
\end{aligned}$$

Therefore, taking limit on T, as T approaches infinity and choosing $x = \log\left(I_T T\right)/(I_T\left(\mu_1-\mu_2\right)^2)$

$$\begin{aligned}
&\lim_{T\to\infty} \mathbb{P}\left(\tau^i_{\pi_{\{1,2\}},1}(T/2) < x\right) \\
&> \lim_{T\to\infty} \mathbb{P}\left(L^i_{\pi,1}\left(\log\left(I_T T\right)/I_T\left(\mu_1-\mu_2\right)^2\right) < \left(\mu_2-y\right)\right)\mathbb{P}\left(L^i_{\pi,2}\left(T/2-\log\left(I_T T\right)/I_T\left(\mu_1-\mu_2\right)^2\right) > \left(\mu_2-y\right)\right) \\
&= e^{2(\mu_2-y)(\mu_1-\mu_2+y)}(1-e^{2y(\mu_2-y)}) = e^c > 0 \text{ for some } c \in (-\infty, 0).
\end{aligned}$$

The last equality is obtained from using Equation 10 and taking the limit. This implies that

$$\begin{aligned}
&\lim_{T\to\infty} \mathbb{P}\left(\tau^f_{\pi_{\{1,2\}},1}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2\right) \\
&> \prod_{i\in\mathbb{I}} \lim_{T\to\infty} \mathbb{P}\left(\tau^i_{\pi_{\{1,2\}},1}(T/2) < \log\left(I_T T\right)/\left(I_T\left(\mu_1-\mu_2\right)^2\right)\right) \\
&= e^{cI_T} = T^{cI_T/\log T}.
\end{aligned}$$

This implies that $\lim_{T\to\infty} \mathbb{P}\left(\tau^f_{\pi,2}(T) > \Omega((I_T T)^a)\right) > T^{cI_T/\log T}$. This implies that

$$Z_\pi(T) > \Omega((I_T T)^a) T^{cI_T/\log T} = \Omega((I_T)^a T^{a+cI_T/\log T}) = \Omega((I_T T)^a)$$

because $\lim_{T\to\infty} I_T/\log T = 0$. This implies that $Z_\pi(T) > \omega((I_T T)^a)$ for all $a \in (0,1)$ proving the claim. □

A.3.3 Proof of Lemma 9

Proof of Lemma 9. We first show that the probability that $\min_{0<\mathbf{1}^T \boldsymbol{s}_j \le IT} C^+_j\left(\boldsymbol{s}_j\right) < \mu_j$ is at most $\frac{\sqrt{2\pi}(\log_4 IT+2)}{JIT}$. By symmetry, the probability that $\max_{0<\mathbf{1}^T \boldsymbol{s}_j \le IT} C^-_j\left(\boldsymbol{s}_j\right) > \mu_j$ is at most $\frac{\sqrt{2\pi}(\log_4 IT+2)}{JIT}$. From the definition,

$$C^+_j\left(\boldsymbol{s}_j\right) = \frac{R^f_j(\boldsymbol{s}_j)}{\mathbf{1}^T\boldsymbol{s}_j} + \alpha\left(\frac{1}{\mathbf{1}^T\boldsymbol{s}_j} + \frac{1}{\sqrt{\mathbf{1}^T\boldsymbol{s}_j}}\right) = \mu_j + \frac{\sum_{i=1}^I B^i_j\left(s^i_j\right)}{\mathbf{1}^T\boldsymbol{s}_j} + \alpha\left(\frac{1}{\mathbf{1}^T\boldsymbol{s}_j} + \frac{1}{\sqrt{\mathbf{1}^T\boldsymbol{s}_j}}\right).$$

Therefore,

$$C^+_j\left(\boldsymbol{s}_j\right) < \mu_j \Leftrightarrow \frac{\sum_{i=1}^I B^i_j\left(s^i_j\right)}{\mathbf{1}^T\boldsymbol{s}_j} + \alpha\left(\frac{1}{\mathbf{1}^T\boldsymbol{s}_j} + \frac{1}{\sqrt{\mathbf{1}^T\boldsymbol{s}_j}}\right) < 0 \Leftrightarrow \sum_{i=1}^I B^i_j\left(s^i_j\right) + \alpha\left(1+\sqrt{\mathbf{1}^T\boldsymbol{s}_j}\right) < 0.$$

Since, $\sum_{i=1}^I B^i_j\left(s^i_j\right)$ is a Wiener process over $\mathbf{1}^T\boldsymbol{s}_j$, the probability

$$\begin{aligned}
&\mathbb{P}\left(\min_{0<\mathbf{1}^T\boldsymbol{s}_j\le IT} C^+_j\left(\boldsymbol{s}_j\right) < \mu_j\right) = \mathbb{P}\left(\min_{0<\mathbf{1}^T\boldsymbol{s}_j\le IT} \sum_{i=1}^I B^i_j\left(s^i_j\right) + \alpha\left(1+\sqrt{\mathbf{1}^T\boldsymbol{s}_j}\right) < 0\right) \\
&= \mathbb{P}\left(\min_{0<s\le IT} B\left(s\right) + \alpha\left(1+\sqrt{s}\right) < 0\right),
\end{aligned}$$

where $B\left(s\right)$ is a standard brownian motion over s. This is the probability of the Brownian motion crossing a square root boundary within a finite time. To obtain this probability, one can try to solve the Fokker-Planck equation of the Volterra type. However, the solution to the resulting differential equation is not known. Instead, we provide an upper bound on this probability by the probability

of a more likely event. We use a piecewise linear function of s that is a lower bound on $\alpha\left(1+\sqrt{s}\right)$. The probability of the Brownian motion crossing this piecewise linear function of s is higher than the probability of the Brownian motion crossing the function$\alpha\left(1+\sqrt{s}\right)$. We note that

$$\begin{aligned}
&\alpha\left(1+\sqrt{s}\right) \geq \alpha\left(1+s\right), \text{ for } 0<s\leq 1 \text{ and}\\
&\alpha\left(1+\sqrt{s}\right) \geq \alpha\left(1+\sqrt{4^n}\right)+\frac{\alpha}{3\sqrt{4^n}}\left(s-4^n\right), \text{ for all } n\geq 0, 4^n\leq s\leq 4^{n+1}.
\end{aligned}$$

This implies that $\alpha\left(1+\sqrt{s}\right) \geq \min\left\{\alpha\left(1+s\right), \min_{n\geq 0}\left(\alpha\left(1+\sqrt{4^n}\right)+\frac{\alpha}{3\sqrt{4^n}}\left(s-4^n\right)\right)\right\}$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\min_{0<s\leq IT} B\left(s\right)+\alpha\left(1+\sqrt{s}\right)<0\right) = \mathbb{P}\left(\max_{0<s\leq IT} B\left(s\right)-\alpha\left(1+\sqrt{s}\right)>0\right)\\
&<\mathbb{P}\left(\max_{0<s\leq IT} B\left(s\right)-\min\left\{\alpha\left(1+s\right), \min_{0\leq n\leq \log_4 IT+1}\left(\alpha\left(1+\sqrt{4^n}\right)+\frac{\alpha}{3\sqrt{4^n}}\left(s-4^n\right)\right)\right\}>0\right)
\end{aligned}$$

We define events

$$\begin{aligned}
A &:= \max_{0<s\leq IT} B\left(s\right)-\alpha\left(1+s\right)>0,\\
A_n &:= \max_{0<s\leq IT} B\left(s\right)-\left(\alpha\left(1+\sqrt{4^n}\right)+\frac{\alpha}{3\sqrt{4^n}}\left(s-4^n\right)\right)>0, \text{ for } n\geq 0.
\end{aligned}$$

Therefore, the above probability,

$$\begin{aligned}
&\mathbb{P}\left(\max_{0<s\leq IT} B\left(s\right)-\min\left\{\alpha\left(1+s\right), \min_{0\leq n\leq \log_4 IT+1}\left(\alpha\left(1+\sqrt{4^n}\right)+\frac{\alpha}{3\sqrt{4^n}}\left(s-4^n\right)\right)\right\}>0\right)\\
&=\mathbb{P}\left(A\cup\left(\cup_{n=0}^{\log_4 IT+1} A_n\right)\right)\\
&<\mathbb{P}\left(A\right)+\sum_{n=0}^{\log_4 IT+1}\mathbb{P}\left(A_n\right)
\end{aligned}$$

We note that

$$\begin{aligned}
&A\subset \hat{A} := \max_{s>0} B\left(s\right)-\alpha\left(1+s\right)>0 \text{ and}\\
&A_n\subset \hat{A}_n := \max_{s>0} B\left(s\right)-\left(\alpha\left(1+\sqrt{4^n}\right)+\frac{\alpha}{3\sqrt{4^n}}\left(s-4^n\right)\right)>0.
\end{aligned}$$

Using the probability of Brownian motion crossing a line[1], we find that $\mathbb{P}\left(\hat{A}\right) = e^{-2\alpha^2}$ and

[1]Probability that the Brownian motion, $B\left(s\right)$ crosses a line $a+bs$ for $a,b>0$ is e^{-2ab}.

$\mathbb{P}\left(\hat{A}_n\right) = e^{-\left(\frac{4}{9}+\frac{2}{3\sqrt{4^n}}\right)\alpha^2}$ for $n \geq 0$. Therefore,

$$\begin{aligned}
&\mathbb{P}(A) + \sum_{n=0}^{\log_4 IT+1} \mathbb{P}(A_n) \\
&< \mathbb{P}\left(\hat{A}\right) + \sum_{n=0}^{\log_4 IT+1} \mathbb{P}\left(\hat{A}_n\right) = e^{-2\alpha^2} + \sum_{n=0}^{\log_4 IT+1} e^{-\left(\frac{4}{9}+\frac{2}{3\sqrt{4^n}}\right)\alpha^2} \\
&< e^{-2\alpha^2} + \sum_{n=0}^{\log_4 IT+1} e^{-\frac{4}{9}\alpha^2} = e^{-2\alpha^2} + (\log_4 IT + 1)\, e^{-\frac{4}{9}\alpha^2} \\
&< (\log_4 IT + 2)\, e^{-\frac{4}{9}\alpha^2}.
\end{aligned}$$

Therefore,

$$\mathbb{P}\left(\min_{0<\mathbf{1}^T \boldsymbol{s}_j \leq IT} C_j^+(\boldsymbol{s}_j) < \mu_j\right) < (\log_4 IT + 2)\, e^{-\frac{4}{9}\alpha^2} = \frac{\sqrt{2\pi}\,(\log_4 IT + 2)}{JIT}.$$

By symmetry,

$$\mathbb{P}\left(\max_{0<\mathbf{1}^T \boldsymbol{s}_j \leq IT} C_j^-(\boldsymbol{s}_j) > \mu_j\right) < (\log_4 IT + 2)\, e^{-\frac{4}{9}\alpha^2} = \frac{\sqrt{2\pi}\,(\log_4 IT + 2)}{JIT}.$$

□

A.3.4 Proof of Lemma 10

Proof of Lemma 10. Pick $\boldsymbol{s}_j$ such that $\mathbf{1}^T \boldsymbol{s}_j > \frac{4\alpha^2}{\delta^2}$. Therefore,

$$C_j^+(\boldsymbol{s}_j) - C_j^-(\boldsymbol{s}_j) = 2\alpha\left(\frac{1}{\mathbf{1}^T \boldsymbol{s}_j} + \frac{1}{\sqrt{\mathbf{1}^T \boldsymbol{s}_j}}\right) = \frac{\delta^2}{2\alpha} + \delta < 2\delta.$$

Since A_j and B_j are false therefore

$$C_j^+(\boldsymbol{s}_j) - \mu_j < C_j^+(\boldsymbol{s}_j) - C_j^-(\boldsymbol{s}_j) < 2\delta \text{ and}$$

$$\mu_j - C_j^-(\boldsymbol{s}_j) < C_j^+(\boldsymbol{s}_j) - C_j^-(\boldsymbol{s}_j) < 2\delta.$$

□

A.3.5 Proof of Theorem 10

We now provide the proof of Theorem 10. We first need to introduce some definitions and some intermediate results. For simplicity of exposition, we assume that $\log_2 J$ is an integer. Without loss of generality, we assume arms are indexed in the decreasing order of their drifts, i.e. $\mu_1 > \mu_2 > \cdots > \mu_J$. We first define dummy policies π^n for each $n \in \{0, \ldots, \log_2 J\}$ that follow policy π^* but never discard the top 2^n arms, i.e.- never discard the arms in $K_n = \{1, \ldots, 2^n\}$.

Definition 3. Dummy policy π^n*: Discard arm* j *at time* $t > 0$ *if* $C_j^+\left(\boldsymbol{\tau}_{\pi^n,j}(t)\right) \leq l^*_{\pi^n}(t)$ *and* $j \notin K_n$.

We note that the policy π^n behaves as policy π^* as long as $C_j^+\left(\boldsymbol{\tau}_{\pi^n,j}(s)\right) > l^*_{\pi^n}(s)$ for all $j \in \{\{1\}, \ldots, \{2^n\}\}$, for all $s < t$ i,e,- $\mathbb{J}_{\pi^n,t|} = \mathbb{J}_{\pi^*,t}$ and $\boldsymbol{\tau}^i_{j|\pi^n}(t) = \boldsymbol{\tau}^i_{j|\pi^*}(t)$ for all $i \in I$ and $j \in K_n$ given $C_j^+\left(\boldsymbol{\tau}_{\pi^n,j}(s)\right) > l^*_{\pi^n}(s)$ for all $s < t$ and $j \in K_n$.

We now define two sets of events, the first set of events is characterized by the efforts on arms at a given time and the second set of events is characterized by the anchor rate at a given time. The first set of events is defined below.

Definition 4. *We define joint events* $\gamma^\circ_{\pi,n}(t)$ *at any time* t*, under policy* π *as an intersection of two different kinds of events for each* n *with* $0 \leq n \leq \log_2 J$ *as following:*

1. $\gamma^1_{\pi,n}(t)$*: at time* t*, under policy* π *all sets of* x *agents, for all* $\frac{I}{3} \leq x \leq I$ *spent a combined cumulative effort of at least* $3\left(\frac{x}{I} - \frac{1}{3}\right)2^n \frac{16\alpha^2}{\Delta^2}$ *on the arms* $K_n \subseteq \mathbb{J}$ *i.e.-*

$$\gamma^1_{\pi,n}(t) \equiv \{\omega \in \Omega : \sum_{i \in G, j \in K_n} \tau^i_{\pi,j}(t) \geq 3\left(\frac{|G|}{I} - \frac{1}{3}\right)2^n \frac{16\alpha^2}{\Delta^2},\ \textit{for all } G \subseteq \mathbb{I} \textit{ with } |G| \geq \frac{I}{3}\}.$$

2. $\gamma^2_{\pi,n}(t)$*: at time* t*, under policy* π *each arm in* K_{n-1} *has total cumulative effort less than* $\frac{16\alpha^2}{\Delta^2}$ *by time* t *i.e.-*

$$\gamma^2_{\pi,n}(t) \equiv \{\omega \in \Omega : \tau^f_{\pi,j}(t) < \frac{16\alpha^2}{\Delta^2} \textit{ for all } j \in K_{n-1}\}.$$

$$\gamma^\circ_{\pi,n}(t) = \gamma^1_{\pi,n}(t) \cap \gamma^2_{\pi,n}(t).$$

The second set of events is defined below.

Definition 5. *We define* $\gamma^3_{\pi,n}(t)$ *as the event that under policy* π *the anchor rate at time* t *is lower than the lowest upper confidence bound for all arms in* K_{n-1} *until time* t*, i.e.-* $l^*_\pi(t) < \inf_{0<s\leq t} C_j^+\left(\boldsymbol{\tau}_{\pi,j}(s)\right)$ *for all* $j \in K_{n-1}$.

For the following, we will use a notation

$$t^\circ_n = \begin{cases} \frac{32\alpha^2 J}{\Delta^2 I},\ \text{if } n = \log_2 J \\ \frac{112\alpha^2 J}{\Delta^2 I}\left(1 - \frac{2^n}{J}\right),\ \text{if } n \in \{1, \ldots, \log_2 J - 1\} \end{cases} \tag{31}$$

We note two points. First, $t_1^\circ < \frac{112\alpha^2}{\Delta^2 I}(J-1)$ for all $J = 2^m$, where $m \in \mathbb{Z}^+$. Second, t_n° clearly depends upon I, J, and T but we are suppressing the notation for the simplicity of notation. We now provide two lemmas needed for the proof of Theorem 10. The proofs are provided following the proof of Theorem 10.

Lemma 13. *Under assumption 1 and under policy π^*, if none of the events A_j, B_j occur, for all $j \in \mathbb{J}$ and none of the events $\gamma^\circ_{\pi^*,n}(t_n^\circ) \cap \gamma^3_{\pi^*,n}(t_n^\circ)$ occur for all $1 \leq n \leq \log_2 J$, then at time t_n°, the anchor rate $l^*_{\pi^*}(t_n^\circ) \geq \mu_{2^{n-1}} - \Delta$ for all $1 \leq n \leq \log_2 J$.*

Lemma 14. *Under assumption 1, $\mathbb{P}\left(\gamma^\circ_{\pi^*,n}(t_n^\circ) \cap \gamma^3_{\pi^*,n}(t_n^\circ)\right) < e^{-\frac{I}{12}}$, for $1 \leq n \leq \log_2 J - 1$.*

We now give the proof of theorem 10.

Proof of Theorem 10. By Lemma 13, if none of the events A_j, B_j occur, for all $j \in \mathbb{J}$ and none of the events $\gamma^\circ_{\pi^*,n}(t_n^\circ) \cap \gamma^3_{\pi^*,n}(t_n^\circ)$ occur for $1 \leq n \leq \log_2 J$ respectively then under policy π^* the anchor rates $l^*_{\pi^*}(t_1^\circ) \geq \mu_1 - \Delta$ and therefore, $l^*_{\pi^*}\left(\frac{112\alpha^2}{\Delta^2 I}(J-1)\right) \geq \mu_1 - \Delta$. The probability that none of the events in A_j, B_j for all $j \in \mathbb{J}$ and none of the events in $\gamma^\circ_{\pi^*,n}(t_n^\circ) \cap \gamma^3_{\pi^*,n}(t_n^\circ)$ occur for $1 \leq n \leq \log_2 J$ is

$$1 - \mathbb{P}\left(\cup_{j\in\mathbb{J}} A_j \cup B_j \cup_{1\leq n\leq \log_2 J} \left(\gamma^\circ_{\pi^*,n}(t_n^\circ) \cap \gamma^3_{\pi^*,n}(t_n^\circ)\right)\right)$$
$$\geq 1 - \sum_{j\in\mathbb{J}} \mathbb{P}(A_j) - \sum_{j\in\mathbb{J}} \mathbb{P}(B_j) - \sum_{1\leq n\leq \log_2 J} \mathbb{P}\left(\gamma^\circ_{\pi^*,n}(t_n^\circ) \cap \gamma^3_{\pi^*,n}(t_n^\circ)\right)$$

which by Lemmas 9 and 14 is at least

$$1 - J\frac{2\sqrt{2\pi}\,(\log_4 IT + 2)}{JIT} - \log_2 J e^{-\frac{I}{12}} = 1 - \frac{2\sqrt{2\pi}\,(\log_4 IT + 2)}{IT} - \log_2 J e^{-\frac{I}{12}}.$$

This completes the proof of the Theorem. □

We now give the proofs of Lemmas 13 and 14.

Proof of Lemma 13. We first show that if A_j, B_j are false for all $j \in \mathbb{J}$ and $\gamma^\circ_{\pi^*,\log_2 J}(t^\circ_{\log_2 J}) \cap \gamma^3_{\pi^*,\log_2 J}(t^\circ_{\log_2 J})$ is false then the anchor rate $l^*_{\pi^*}(t^\circ_{\log_2 J}) \geq \mu_{\frac{J}{2}} - \Delta$. Assume that A_j, B_j are false for all $j \in \mathbb{J}$ and $\gamma^\circ_{\pi^*,\log_2 J}(t^\circ_{\log_2 J}) \cap \gamma^3_{\pi^*,\log_2 J}(t^\circ_{\log_2 J})$ is false. We note that $\gamma^1_{\pi^*,\log_2 J}(t^\circ_{\log_2 J})$ is true because

$$\sum_{i\in G, j\in\mathbb{J}} \tau^i_{\pi^*,j}(t^\circ_{\log_2 J}) = 2\frac{|G|}{I} \times \frac{16\alpha^2 J}{\Delta^2} \geq 3\left(\frac{|G|}{I} - \frac{1}{3}\right)\frac{16\alpha^2 J}{\Delta^2} \text{ for all } G \subseteq \mathbb{I} \text{ with } \frac{I}{3} \leq |G| \leq I.$$

This implies that $\gamma^2_{\pi^*,\log_2 J}(t^\circ_{\log_2 J})$ or $\gamma^3_{\pi^*,\log_2 J}(t^\circ_{\log_2 J})$ must be false. Therefore, either there exists a product $j_1 \in K_{\log_2 J - 1}$ such that $\tau^f_{\pi^*,j_1}(t^\circ_{\log_2 J}) \geq \frac{16\alpha^2}{\Delta^2}$ or there exists a product $j_2 \in K_{\log_2 J - 1}$ such that $l^*_{\pi^*}(t^\circ_{\log_2 J}) \geq \inf_{0<s\leq t^\circ_{\log_2 J}} C^+_{j_2}(\boldsymbol{\tau}_{\pi^*,j_2}(s))$. Since $A_{j_1}, A_{j_2}, B_{j_1}, B_{j_2}$ are all false, therefore

either $l^*_{\pi^*}(t^\circ_{\log_2 J}) \geq \sup_{0<s\leq t^\circ_{\log_2 J}} C^-_{j_1}\left(\boldsymbol{\tau}_{\pi^*,j_1}(s)\right) \geq \mu_{j_1} - \Delta \geq \mu_{\frac{J}{2}} - \Delta$ by lemma 10 or $l^*_{\pi^*}(t^\circ_{\log_2 J}) \geq \inf_{0<s\leq t^\circ_{\log_2 J}} C^+_{j_2}\left(\boldsymbol{\tau}_{\pi^*,j_2}(s)\right) \geq \mu_{j_2} > \mu_{\frac{J}{2}} - \Delta$.

We will now show that under the assumptions of the Lemma, at time t°_n, the anchor rate $l^*_{\pi^*}(t^\circ_n) \geq \mu_{2^{n-1}} - \Delta$ for all $n \in \{1 \dots \log_2 J - 1\}$. We define t_n as the first time $l^*_{\pi^*}(t)$ hits $\mu_{2^{n-1}} - \Delta$, i.e.-

$$t_n = \inf\{t | l^*_{\pi^*}(t) \geq \mu_{2^{n-1}} - \Delta\}.$$

We point out that by definition, t_n is decreasing in n. We also define x^n_j as the total effort by all agents on product j between times t_{n+1} and t_n, i.e.-

$$x^n_j = \tau^f_{\pi^*,j}\left(t_n\right) - \tau^f_{\pi^*,j}\left(t_{n+1}\right).$$

Since A_j, B_j are false for all $j \in \mathbb{J}$ therefore by Lemma 10 for all $j \in K_{m+1} \setminus K_m$, $0 \leq m \leq \log_2 J - 1$, $C^+_j\left(\boldsymbol{\tau}_{\pi^*,j}(t)\right) \leq \mu_j + \Delta \leq l^*_{|\pi^*}(t)$ for $t \geq t_{m+1}$ if $\tau^f_{\pi^*,j}(t) \geq \frac{16\alpha^2}{\Delta^2}$. Pick any $j \in K_{m+1} \setminus K_m$. The product j would be discarded under policy π^* if it has the total cumulative effort of at least $\frac{16\alpha^2}{\Delta^2}$ at any time $t \geq t_{m+1}$. This implies that $\tau^f_{\pi^*,j}(t) \leq \max\{\tau^f_{\pi^*,j}\left(t_{m+1}\right), \frac{16\alpha^2}{\Delta^2}\}$, for all $t \geq t_m$. This implies that

$$\sum_{n=0}^{m} x^n_j \leq \frac{16\alpha^2}{\Delta^2}.$$

We will first show that $t_n - t_{n+1} \leq \frac{1}{I}\max\{3\sum_{j\notin K_n} x^n_j, \sum_{j\notin K_n} x^n_j + 2^{n+1}\frac{16\alpha^2}{\Delta^2}\}$. Then using the principle of optimality and mathematical induction, we will prove the lemma. At $t = t_{n+1} + \frac{1}{I}\max\{3\sum_{j\notin K_n} x^n_j, \sum_{j\notin K_n} x^n_j + 2^{n+1}\frac{16\alpha^2}{\Delta^2}\}$, $\gamma^1_{\pi^*,n}(t)$ is true because for any set of agents $G \subseteq \mathbb{I}$,

$$\begin{aligned}
&\sum_{i\in G, j\in K_n} \tau^i_{\pi^*,j}(t) \\
&\geq \sum_{i\in G, j\in K_n} \tau^i_{\pi^*,j}(t) - \tau^i_{\pi^*,j}(t_{n+1}) \\
&\geq |G|(t - t_{n+1}) - \sum_{j\notin K_n} x^n_j \\
&\geq \frac{|G|}{I}\max\{3\sum_{j\notin K_n} x^n_j, \sum_{j\notin K_n} x^n_j + 2^{n+1}\frac{16\alpha^2}{\Delta^2}\} - \sum_{j\notin K_n} x^n_j
\end{aligned}$$

If $\sum_{j\notin K_n} x_j^n > 2^n \frac{16\alpha^2}{\Delta^2}$ then

$$\begin{aligned}
&\frac{|G|}{I}\max\{3\sum_{j\notin K_{n+1}} x_j^n, \sum_{j\notin K_n} x_j^n + 2^{n+1}\frac{16\alpha^2}{\Delta^2}\} - \sum_{j\notin K_n} x_j^n \\
&= 3\frac{|G|}{I}\sum_{j\notin K_n} x_j^n - \sum_{j\notin K_n} x_j^n \\
&> 3(\frac{|G|}{I} - \frac{1}{3})2^{n+1}\frac{16\alpha^2}{\Delta^2}
\end{aligned}$$

If $\sum_{j\notin K_{n+1}} x_j^n \leq 2^n \frac{16\alpha^2}{\Delta^2}$ then

$$\begin{aligned}
&\frac{|G|}{I}\max\{3\sum_{j\notin K_n} x_j^n, \sum_{j\notin K_n} x_j^n + 2^{n+1}\frac{16\alpha^2}{\Delta^2}\} - \sum_{j\notin K_n} x_j^n \\
&= \frac{|G|}{I}(\sum_{j\notin K_n} x_j^n + 2^{n+1}\frac{16\alpha^2}{\Delta^2}) - \sum_{j\notin K_n} x_j^n
\end{aligned}$$

$$\begin{aligned}
&= \frac{|G|}{I}2^{n+1}\frac{16\alpha^2}{\Delta^2} - (1-\frac{|G|}{I})\sum_{j\notin K_n} x_j^n \\
&\geq \frac{|G|}{I}2^{n+1}\frac{16\alpha^2}{\Delta^2} - (1-\frac{|G|}{I})2^n\frac{16\alpha^2}{\Delta^2} \\
&= 3(\frac{|G|}{I} - \frac{1}{3})2^n\frac{16\alpha^2}{\Delta^2}
\end{aligned}$$

This implies that $\gamma^2_{\pi^*,n}(t)$ or $\gamma^3_{\pi^*,n}(t)$ must be false. Therefore, there exists a product $j_1 \in K_{n-1}$ such that $\tau^f_{\pi^*,j_1}(t) \geq \frac{16\alpha^2}{\Delta^2}$ or there exists a product $j_2 \in K_{n-1}$ such that $l^*_{\pi^*}(t) \geq \inf_{0<s\leq t} C^+_{j_2}(\boldsymbol{\tau}_{\pi^*,j_2}(s))$. Since $A_{j_1}, A_{j_2}, B_{j_1}, B_{j_2}$ are all false, therefore by lemma 10 either $l^*_{\pi^*}(t) \geq \sup_{0<s\leq t} C^-_{j_1}(\boldsymbol{\tau}_{\pi^*,j_1}(s)) \geq \mu_{j_1} - \Delta \geq \mu_{2^n-1} - \Delta$ or $l^*_{\pi^*}(t) \geq \inf_{0<s<t} C^+_{j_2}(\boldsymbol{\tau}_{\pi^*,j_2}(s)) \geq \mu_{j_2} > \mu_{2^{n-1}} - \Delta$. Therefore,

$$t_n \leq t = t_{n+1} + \frac{1}{I}\max\{3\sum_{j\notin K_n} x_j^n, \sum_{j\notin K_n} x_j^n + 2^{n+1}\frac{16\alpha^2}{\Delta^2}\}.$$

We will now complete the proof of the lemma using the principle of optimality and mathematical induction. For each $1 < m \leq \log_2 J$ and $1 \leq n < m$, we define y_j^n as the total cumulative effort by all agents on product $j \in K_m \setminus K_{m-1}$ between times t_n and t_m, i.e.-

$$y_j^n = \sum_{l=n}^{m-1} x_j^l = \tau^f_{\pi^*,j}(t_n) - \tau^f_{\pi^*,j}(t_m),$$

$\boldsymbol{y}^n = \{y_j^n\}_{j \notin K_n}$ is a vector whose elements are y_j^n for $j \notin K_n$.

We define $t_n^*(\boldsymbol{y}^n)$ as the maximum possible value of t_n under the conditions of the Lemma and the policy π^* given $\boldsymbol{y}^n$. When $J = 2$ then clearly $t_n \leq t^*(1)(0) = t_1^\circ$. Assume $J > 2$. For $n = \log_2 J - 1$,

$$\begin{aligned} t_n^*(\boldsymbol{y}^n) &\leq \frac{32\alpha^2 J}{\Delta^2 I} + \frac{1}{I}\max\{3\mathbf{1}^T\boldsymbol{y}^n, \mathbf{1}^T\boldsymbol{y}^n + J\frac{16\alpha^2}{\Delta^2}\} \\ &= \begin{cases} \frac{32\alpha^2 J}{\Delta^2 I} + \frac{1}{I}\mathbf{1}^T\boldsymbol{y}^n + \frac{16\alpha^2 J}{\Delta^2 I}, & \text{if } \mathbf{1}^T\boldsymbol{y}^n < J\frac{8\alpha^2}{\Delta^2} \\ \frac{56\alpha^2 J}{\Delta^2 I}, & \text{if } \mathbf{1}^T\boldsymbol{y}^n = J\frac{8\alpha^2}{\Delta^2} \end{cases} \end{aligned}$$

because $\mathbf{1}^T\boldsymbol{y}^n \leq J\frac{8\alpha^2}{\Delta^2}$. This implies that for $n = \log_2 J - 1$, since $\mathbf{1}^T\boldsymbol{y}^n \leq J\frac{8\alpha^2}{\Delta^2}$ therefore $t_n^*(\boldsymbol{y}^n) \leq \frac{56\alpha^2 J}{\Delta^2 I} = \frac{112\alpha^2 J}{\Delta^2 I}(1 - \frac{2^n}{J})$. We now state the assumption for induction. Assume that for some $n \leq \log_2 J - 2$,

$$t_{n+1}^*(\boldsymbol{y}^{n+1}) \leq \begin{cases} \frac{64\alpha^2 J}{\Delta^2 I}(1 - \frac{2^n}{J}) + \frac{1}{I}\mathbf{1}^T\boldsymbol{y}^{n+1}, & \text{if } \mathbf{1}^T\boldsymbol{y}^{n+1} < 2^{n+1}\frac{16\alpha^2}{\Delta^2} \\ \frac{64\alpha^2 J}{\Delta^2 I}(1 - \frac{2^{n+1}}{J}) + \frac{3}{I}\mathbf{1}^T\boldsymbol{y}^{n+1}, & \text{if } \mathbf{1}^T\boldsymbol{y}^{n+1} \geq 2^{n+1}\frac{16\alpha^2}{\Delta^2} \end{cases}$$

Clearly, this holds for $n = \log_2 J - 2$. Therefore using the principle of optimality,

$$t_n^*(\boldsymbol{y}^n) = \max_{\{x_j^n \leq y_j^n\}_{j \notin K_{n+1}}} \{t_{n+1}^*(\boldsymbol{y}^{n+1}) + \frac{1}{I}\max\{3\sum_{j \notin K_n} x_j^n, \sum_{j \notin K_n} x_j^n + 2^{n+1}\frac{16\alpha^2}{\Delta^2}\}\},$$

where $y_j^{n+1} = y_j^n - x_j^n$ for all $j \notin K_{n+1}$. Since, $\mathbf{1}^T\boldsymbol{y}^n \geq \mathbf{1}^T\boldsymbol{y}^{n+1}$, therefore $\mathbf{1}^T\boldsymbol{y}^{n+1} \geq 2^{n+1}\frac{16\alpha^2}{\Delta^2}$ implies $\mathbf{1}^T\boldsymbol{y}^n \geq 2^n\frac{16\alpha^2}{\Delta^2}$. This implies that

$$\begin{aligned} t_n^*(\boldsymbol{y}^n) &= t_{n+1}^*(0) + \frac{1}{I}\max\{3\mathbf{1}^T\boldsymbol{y}^n, \mathbf{1}^T\boldsymbol{y}^n + 2^{n+1}\frac{16\alpha^2}{\Delta^2}\} \\ &\leq \begin{cases} \frac{64\alpha^2 J}{\Delta^2 I}(1 - \frac{2^n}{J}) + \frac{1}{I}\mathbf{1}^T\boldsymbol{y}^n + 2^{n+1}\frac{16\alpha^2}{\Delta^2 I} = \frac{64\alpha^2 J}{\Delta^2 I}(1 - \frac{2^{n-1}}{J}) + \mathbf{1}^T\boldsymbol{y}^n, & \text{if } \mathbf{1}^T\boldsymbol{y}^n < 2^n\frac{16\alpha^2}{\Delta^2} \\ \frac{64\alpha^2 J}{\Delta^2 I}(1 - \frac{2^n}{J}) + \frac{3}{I}\mathbf{1}^T\boldsymbol{y}^n, & \text{if } \mathbf{1}^T\boldsymbol{y}^n \geq 2^n\frac{16\alpha^2}{\Delta^2} \end{cases} \end{aligned}$$

By induction, the above holds for all $n \in \{1, \ldots, \log_2 J - 2\}$. The maximum value is achieved when $y_j^n = \frac{16\alpha^2}{\Delta^2}$ for all $j \notin K_n$ because $t_n^*(\boldsymbol{y}^n)$ is increasing in each component of $\boldsymbol{y}^n$. We define $y_*^n = [\frac{16\alpha^2}{\Delta^2}, \ldots, \frac{16\alpha^2}{\Delta^2}]$ a vector of all $\frac{16\alpha^2}{\Delta^2}$. Therefore,

$$\begin{aligned} t_n &\leq t_n^*(y_*^n) \\ &= \frac{64\alpha^2 J}{\Delta^2 I}(1 - \frac{2^n}{J}) + \frac{3J}{I}\frac{16\alpha^2}{\Delta^2}(1 - \frac{2^n}{J}) = \frac{112\alpha^2 J}{\Delta^2 I}(1 - \frac{2^n}{J}) = t_n^\circ. \end{aligned}$$

This completes the proof of the lemma.

□

Proof of Lemma 14. We first point that the events $\gamma^{\circ}_{\pi^*,n}(t^{\circ}_n) \cap \gamma^{3}_{\pi^*,n}(t^{\circ}_n)$ and $\gamma^{\circ}_{\pi^{n-1},n}(t^{\circ}_n) \cap \gamma^{3}_{\pi^{n-1},n}(t^{\circ}_n)$ for all $1 \leq n \leq \log_2 J$ are equivalent, i.e.- $\gamma^{\circ}_{\pi^*,n}(t^{\circ}_n) \cap \gamma^{3}_{\pi^*,n}(t^{\circ}_n) \Leftrightarrow \gamma^{\circ}_{\pi^{n-1},n}(t^{\circ}_n) \cap \gamma^{3}_{\pi^{n-1},n}(t^{\circ}_n)$. This is because if $l^*_{\pi^*}(t) < \inf_{0<s\leq t} C^+_j(\boldsymbol{\tau}_{\pi^*,j}(s))$ for all $j \in K_{n-1}$, then the firm never discards any of the products in K_{n-1} until time t°_n under policy π^*. Therefore the firm and all agents take the same actions until time t°_n under policy π^{n-1} as they would under policy π^*. This implies that $\tau^i_{\pi^*,j}(s) = \tau^i_{\pi^{n-1},j}(s)$ for all $0 < s \leq t^{\circ}_n$ and all agents $i \in \mathbb{I}$ and therefore $\gamma^{\circ}_{\pi^*,n}(t^{\circ}_n) \cap \gamma^{3}_{\pi^*,n}(t^{\circ}_n) \Leftrightarrow \gamma^{\circ}_{\pi^{n-1},n}(t^{\circ}_n) \cap \gamma^{3}_{\pi^{n-1},n}(t^{\circ}_n)$. This implies that

$$\mathbb{P}\left(\gamma^{\circ}_{\pi^*,n}(t^{\circ}_n) \cap \gamma^{3}_{\pi^*,n}(t^{\circ}_n)\right) = \mathbb{P}\left(\gamma^{\circ}_{\pi^{n-1},n}(t^{\circ}_n) \cap \gamma^{3}_{\pi^{n-1},n}(t^{\circ}_n)\right) \leq \mathbb{P}\left(\gamma^{\circ}_{\pi^{n-1},n}(t^{\circ}_n)\right).$$

Further,

$$\mathbb{P}\left(\gamma^{\circ}_{\pi^{n-1},n}(t^{\circ}_n)\right) = \mathbb{P}\left(\gamma^{1}_{\pi^{n-1},n}(t^{\circ}_n) \cap \gamma^{2}_{\pi^{n-1},n}(t^{\circ}_n)\right) \leq \mathbb{P}\left(\gamma^{2}_{\pi^{n-1},n}(t^{\circ}_n) | \gamma^{1}_{\pi^{n-1},n}(t^{\circ}_n)\right).$$

We point that

$$\gamma^{2}_{\pi^{n-1},n}(t^{\circ}_n) \Rightarrow \sum_{j\in K_{n-1}} \tau^f_{\pi^{n-1},j}(t^{\circ}_n) < 2^{n-1}\frac{16\alpha^2}{\Delta^2}.$$

We also define the set

$$X_n = \{\boldsymbol{x} \in \mathbb{R}^I : \sum_{i\in G} x_i \geq 3\left(\frac{G}{I} - \frac{1}{3}\right)2^n \frac{16\alpha^2}{\Delta^2}, \text{ for all } G \subseteq I, |G| > \frac{I}{3}\}$$

and vector $\boldsymbol{y}^n \in \mathbb{R}^I$ where the ith element $y^n_i = \sum_{j\in K_n} \tau^i_{\pi^{n-1},j}(t_n)$. Therefore

$$\gamma^{1}_{\pi^{n-1},n}(t^{\circ}_n) \Leftrightarrow \boldsymbol{y}^n \in X_n.$$

Therefore

$$\mathbb{P}\left(\gamma^{2}_{\pi^{n-1},n}(t^{\circ}_n) | \gamma^{1}_{\pi^{n-1},n}(t^{\circ}_n)\right) \leq \mathbb{P}\left(\sum_{j\in K_{n-1}} \tau^f_{\pi^{n-1},j}(t) < 2^{n-1}\frac{16\alpha^2}{\Delta^2} | \boldsymbol{y}^n \in X_n\right).$$

We next point out that by Theorem 6 and Lemma 12,

$$\boldsymbol{E}\left[\sum_{j\in K_{n-1}} \tau^i_{\pi^{n-1},j}(t) \,|y^n_i = x_i\right] \geq \boldsymbol{E}\left[\sum_{j\in K_{n-1}} \tau^i_{\pi_{K_n},j}(x_i)\right] > \frac{x_i}{2}.$$

By Azuma-Hoeffding inequality,

$$\mathbb{P}\left(\sum_{i\in I, j\in K_{n-1}} \tau^i_{\pi^{n-1},j}(t_n) < 2^{n-1}\frac{16\alpha^2}{\Delta^2} | \boldsymbol{y}^n = \boldsymbol{x}\right) < e^{-2\frac{(2^{n-1}\frac{16\alpha^2}{\Delta^2} - \frac{\boldsymbol{x}.\mathbf{1}}{2})^2}{\boldsymbol{x}.\boldsymbol{x}}}.$$

The maximum value of the right hand side of the inequality over all $\boldsymbol{x} \in X_n$ is attained at $\boldsymbol{x}^*$, where

$$x^*_i = \begin{cases} 0, & \text{if } i \le \frac{I}{3} \\ 3.2^n \frac{16\alpha^2}{I\Delta^2}, & \text{if } i > \frac{I}{3} \end{cases}$$

and the maximum value is $e^{-2I\frac{(2^{n-1}\frac{16\alpha^2}{\Delta^2})^2}{6(2^n\frac{16\alpha^2}{\Delta^2})^2}} = e^{-\frac{I}{12}}$. Therefore for each $n \in \{1, \ldots, \log_2 J\}$,

$$\mathbb{P}\left(\gamma^\circ_{\pi^*,n}(t^\circ_n) \cap \gamma^3_{\pi^*,n}(t^\circ_n)\right) < e^{-\frac{I}{12}}.$$

□

A.3.6 Proof of Theorem 11

Proof. Under assumption 1 by Theorem 10 and Lemma 9, the anchor rate $l^*_{\pi^*}(t) \ge \mu_{\{1\}} - \Delta$ by time $t = \frac{112\alpha^2}{\Delta^2 I}(J-1)$ and A_j, B_j are false for all $j \in J$ with probability at least $1 - \frac{2\sqrt{2\pi}(\log_4 IT+2)}{IT} - \log_2 J e^{-\frac{I}{12}}$. Therefore after time $t = \frac{112\alpha^2}{\Delta^2 I}(J-1)$ all arms other than arm $\{1\}$ that receive the total cumulative effort of at least $\frac{16\alpha^2}{\Delta^2}$ are discarded with probability at least $1 - \frac{2\sqrt{2\pi}(\log_4 IT+2)}{IT} - \log_2 J e^{-\frac{I}{12}}$. Therefore the total cumulative effort on arms other than arm $\{1\}$ is at most $\frac{112\alpha^2}{\Delta^2}(J-1) + \frac{16\alpha^2}{\Delta^2}(J-1)$ with probability at least $1 - \frac{2\sqrt{2\pi}(\log_4 IT+2)}{IT} - \log_2 J e^{-\frac{I}{12}}$. This implies that the expected total cumulative effort on all arms other than arm $\{1\}$ is:

$$\begin{aligned}
&\boldsymbol{E}[\sum_{j\neq\{1\}} \tau^f_{\pi^*,j}(T)] \le \frac{112\alpha^2}{\Delta^2}(J-1) + \frac{16\alpha^2}{\Delta^2}(J-1) + \left(\frac{2\sqrt{2\pi}\,(\log_4 IT+2)}{IT} + \log_2 J e^{-\frac{I}{12}}\right) IT \\
&= \frac{36}{\Delta^2}(8J-15)\left(\log J + \log IT - \log\sqrt{2\pi}\right) + \sqrt{2\pi}\,(\log_2 IT + 4) + \frac{12\beta\log_2 J \log T}{T^{\beta-1}} \\
&< \left(\frac{576J}{\Delta^2} + \frac{\sqrt{2\pi}}{\log 2}\right)\log IT + \frac{12\beta\log_2 J\log T}{T^{\beta-1}}
\end{aligned}$$

where $\beta = \frac{I}{12\log T} > 1$ and $J < IT$. □

A.3.7 Proof of Theorem 12

Proof of Theorem 12. Assume $\mu_1 > \mu_2$. By Lemma 8 any policy π with regret $o(T^a)$ for all $a > 0$ for all $\boldsymbol{\mu}'$ with $\Delta > 0$ must have

$$\lim_{T\to\infty} \mathbb{P}\left(t^2_{\pi,1} < 2\left(1-\epsilon\right)\log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2, t^2_{\pi,2} < \frac{I_T T}{2}\right) = 0$$

for all $\epsilon > 0$. Therefore, we can restrict the candidate policies to the set of policies that satisfy this condition. Asymptotically, such policies do not discard any of the products $2, 3, \ldots, J$ before $T/2$ unless $\tau^f_{\pi,1}(T/2) > 2\left(1-\epsilon\right)\log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2$ almost surely. Such policies also discard product 1 before time $T/2$ with probability at most $o((I_T T)^{a-1})$ for all $0 < a < 1$ otherwise the expected regret for $\boldsymbol{\mu}_T$ would be $\Omega((I_T T)^a)$. This implies that for $0 < a < 1$ by setting $\epsilon = 1/2$

$$\begin{aligned}
&\lim_{T\to\infty} \mathbb{P}\left(T/2 - \tau^f_{\pi,1}(T) > \Omega((I_T T)^a)\right) \\
&> \mathbb{P}\left(T/2 - \tau^f_{\pi,1}(T/2) > \Omega((I_T T)^a), \mathbb{J}_{T/2} = \mathbb{J}\right) + \lim_{T\to\infty} \mathbb{P}\left(1 \notin \mathbb{J}_{T/2}\right) \\
&> \lim_{T\to\infty} \mathbb{P}\left(\tau^f_{\pi,1}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2, \mathbb{J}_{T/2} = \mathbb{J}\right) + \lim_{T\to\infty} \mathbb{P}\left(1 \notin \mathbb{J}_{T/2}\right) \\
&= \lim_{T\to\infty} \mathbb{P}\left(\tau^f_{\pi_{\mathbb{J}},1}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2, 1 \in \mathbb{J}_{T/2}\right) + \lim_{T\to\infty} \mathbb{P}\left(1 \notin \mathbb{J}_{T/2}\right)
\end{aligned}$$

The last equality applies because $\tau^f_{\pi,1}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2$ and $1 \in \mathbb{J}_{T/2}$ almost surely implies that $\mathbb{J}_{T/2|\pi} = \mathbb{J}$ and

$$\tau^f_{\pi,1}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2 \Leftrightarrow \tau^f_{\pi_{\mathbb{J}},1}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2$$

almost surely following Lemma 7. Therefore the above probability is greater than

$$\begin{aligned}
&\lim_{T\to\infty} \mathbb{P}\left(\tau^f_{\pi_{\mathbb{J}},1}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2\right) - \lim_{T\to\infty} \mathbb{P}\left(1 \notin \mathbb{J}_{T/2}\right) + \lim_{T\to\infty} \mathbb{P}\left(1 \notin \mathbb{J}_{T/2}\right) \\
&= \lim_{T\to\infty} \mathbb{P}\left(\tau^f_{\pi_{\mathbb{J}},1}(T/2) < \log\left(I_T T\right)/\left(\mu_1-\mu_2\right)^2\right).
\end{aligned}$$

From Lemma 11 for any agent i and $x < T/2$,

$$\mathbb{P}\left(\tau^i_{\pi_{\mathbb{J}},1}(T/2) < x\right) = \mathbb{P}\left(L^i_{\pi_{\{1\}}}(x) < L^i_{\pi_{\mathbb{J}\setminus\{1\}}}(T/2 - x)\right)$$

We point out that by equation 10 for all $y \leq 0$, $\mathbb{P}\left(L^i_1(x) < y\right)$ is absolutely continuous in μ_1. Therefore,

$$\mathbb{P}\left(L^i_{\pi_{\{1\}}}(x) < L^i_{\pi_{\mathbb{J}\setminus\{1\}}}(T/2 - x)\right)$$

is continuous in μ_1. By taking limit as T approaches infinity and choosing

$$x = \log(I_T T)/(I_T(\mu_1 - \mu_2)^2),$$

x approaches infinity along with T.

$$\begin{aligned}&\lim_{T\to\infty} \mathbb{P}\left(\tau^i_{\pi_{\mathbb{J}},1}(T/2) < x\right)\\ &= \lim_{T\to\infty} \mathbb{P}\left(L^i_{\pi_{\{1\}}}(x) < L^i_{\pi_{\mathbb{J}\setminus\{1\}}}(T/2 - x)\right)\end{aligned}$$

By symmetry, when $\mu_1 = \mu_2$, then the probability

$$\lim_{x,y\to\infty} \mathbb{P}\left(L^i_{\pi_{\{1\}}}(x) < L^i_{\pi_{\mathbb{J}\setminus\{1\}}}(y)\right) = 1 - 1/J.$$

Therefore by continuity, for sufficiently large T, there exists a $\beta > 0$ such that

$$\mathbb{P}\left(L^i_{\pi_{\{1\}}}(x) < L^i_{\pi_{\mathbb{J}\setminus\{1\}}}(T/2 - x)\right) > 1 - 1/(\sqrt{(1-\beta)}J).$$

This implies that

$$\begin{aligned}&\lim_{T\to\infty} \mathbb{P}\left(\tau^f_{\pi_{\{1,2\}},1}(T/2) < \log(I_T T)/(\mu_1 - \mu_2)^2\right)\\ &> \prod_{i\in\mathbb{I}} \lim_{T\to\infty} \mathbb{P}\left(\tau^f_{\pi_{\{1,2\}},1}(T/2) < \log(I_T T)/(I_T(\mu_1 - \mu_2)^2)\right)\\ &> (1 - \frac{1}{\sqrt{(1-\beta)}J})^I_T > e^{-I_T/(\sqrt{(1-\beta)}J)} = T^{-I_T/(\sqrt{(1-\beta)}J\log T)} > T^{-\sqrt{(1-\beta)}}.\end{aligned}$$

This implies that $\lim_{T\to\infty} \mathbb{P}\left(\tau^f_{\pi,2}(T) > \Omega((I_T T)^a)\right) > T^{-\sqrt{(1-\beta)}}$ for all $a \in (0,1)$. This implies that

$$\tilde{Z}_\pi(T, \boldsymbol{\mu}_T) > \Omega((I_T T)^a) T^{-\sqrt{(1-\beta)}} = \Omega((I_T)^a T^{a-\sqrt{(1-\beta)}}).$$

By choosing $a = \sqrt{(1-\beta)} - (1-\beta)$, and $b = 1 - \sqrt{(1-\beta)}$,

$$\tilde{Z}_\pi(T, \boldsymbol{\mu}_T) = \Omega((I_T T)^b).$$

□