

---

# Bandits with Delayed, Aggregated Anonymous Feedback

---

Ciara Pike-Burke<sup>1</sup> Shipra Agrawal<sup>2</sup> Csaba Szepesvári<sup>3,4</sup> Steffen Grünewälder<sup>1</sup>

## Abstract

We study a variant of the stochastic  $K$ -armed bandit problem, which we call “bandits with delayed, aggregated anonymous feedback”. In this problem, when the player pulls an arm, a reward is generated, however it is not immediately observed. Instead, at the end of each round the player observes only the sum of a number of previously generated rewards which happen to arrive in the given round. The rewards are stochastically delayed and due to the aggregated nature of the observations, the information of which arm led to a particular reward is lost. The question is what is the cost of the information loss due to this delayed, aggregated anonymous feedback? Previous works have studied bandits with stochastic, non-anonymous delays and found that the regret increases only by an additive factor relating to the expected delay. In this paper, we show that this additive regret increase can be maintained in the harder delayed, aggregated anonymous feedback setting when the expected delay (or a bound on it) is known. We provide an algorithm that matches the worst case regret of the non-anonymous problem exactly when the delays are bounded, and up to logarithmic factors or an additive variance term for unbounded delays.

## 1. Introduction

The stochastic multi-armed bandit (MAB) problem is a prominent framework for capturing the exploration-exploitation tradeoff in online decision making and experiment design. The MAB problem proceeds in discrete sequential rounds, where in each round, the player pulls one

of the  $K$  possible arms. In the classic stochastic MAB setting, the player immediately observes stochastic feedback from the pulled arm in the form of a ‘reward’ which can be used to improve the decisions in subsequent rounds. One of the main application areas of MABs is in online advertising. Here, the arms correspond to adverts, and the feedback would correspond to *conversions*, that is users buying a product after seeing an advert. However, in practice, these conversions may not necessarily happen immediately after the advert is shown, and it may not always be possible to assign the credit of a sale to a particular showing of an advert. A similar challenge is encountered in many other applications, e.g., in personalized treatment planning, where the effect of a treatment on a patient’s health may be delayed, and it may be difficult to determine which out of several past treatments caused the change in the patient’s health; or, in content design applications, where the effects of multiple changes in the website design on website traffic and footfall may be delayed and difficult to distinguish.

In this paper, we propose a new bandit model to handle online problems with such ‘delayed, aggregated and anonymous’ feedback. In our model, a player interacts with an environment of  $K$  actions (or arms) in a sequential fashion. At each time step the player selects an action which leads to a reward generated at random from the underlying reward distribution. At the same time, a nonnegative random integer-valued delay is also generated i.i.d. from an underlying delay distribution. Denoting this delay by  $\tau \geq 0$  and the index of the current round by  $t$ , the reward generated in round  $t$  will arrive at the end of the  $(t + \tau)$ th round. At the end of each round, the player observes only the *sum* of all the rewards that arrive in that round. Crucially, the player does not know which of the past plays have contributed to this aggregated reward. We call this problem *multi-armed bandits with delayed, aggregated anonymous feedback* (MABDAF). As in the standard MAB problem, in MABDAF, the goal is to maximize the cumulative reward from  $T$  plays of the bandit, or equivalently to minimize the regret. The regret is the total difference between the reward of the optimal action and the actions taken.

If the delays are all zero, the MABDAF problem reduces to the standard (stochastic) MAB problem, which has been studied considerably (e.g., Thompson, 1933; Lai & Robbins, 1985; Auer et al., 2002; Bubeck & Cesa-Bianchi,

---

<sup>1</sup> Department of Mathematics and Statistics, Lancaster University, Lancaster, UK <sup>2</sup> Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, USA <sup>3</sup> DeepMind, London, UK <sup>4</sup> Department of Computing Science, University of Alberta, Edmonton, AB, Canada. Correspondence to: Ciara Pike-Burke <ciara.pikeburke@gmail.com>.

Multi-Armed Bandits (eg. <a href="#">Auer et al. (2002)</a> ) $O(\sqrt{KT \log T})$	Delayed Feedback Bandits (eg. <a href="#">Joulani et al. (2013)</a> ) $O(\sqrt{KT \log T} + K\mathbb{E}[\tau])$	Bandits with Delayed, Aggregated Anonymous Feedback $O(\sqrt{KT \log K} + K\mathbb{E}[\tau])$
$\xrightarrow{\hspace{10em}}$ Difficulty		

Figure 1: The relative difficulties and problem independent regret bounds of the different problems. For MABDAF, our algorithm uses knowledge of  $\mathbb{E}[\tau]$  and a mild assumption on a delay bound, which is not required by [Joulani et al. \(2013\)](#).

2012). Compared to the MAB problem, the job of the player in our problem appears to be significantly more difficult since the player has to deal with (i) that some feedback from the previous pulls may be *missing* due to the delays, and (ii) that the feedback takes the form of the sum of an *unknown number* of rewards of *unknown origin*.

An easier problem is when the observations are delayed, but they are *non-aggregated* and *non-anonymous*: that is, the player has to only deal with challenge (i) and not (ii). Here, the player receives delayed feedback in the shape of action-reward pairs that inform the player of both the individual reward and which action generated it. This problem, which we shall call the (*non-anonymous*) *delayed feedback bandit problem*, has been studied by [Joulani et al. \(2013\)](#), and later followed up by [Mandel et al. \(2015\)](#) (for bounded delays). Remarkably, they show that compared to the standard (non-delayed) stochastic MAB setting, the regret will increase only additively by a factor that scales with the expected delay. For delay distributions with a finite expected delay,  $\mathbb{E}[\tau]$ , the worst case regret scales with  $O(\sqrt{KT \log T} + K\mathbb{E}[\tau])$ . Hence, the price to pay for the delay in receiving the observations is negligible. QPM-D of [Joulani et al. \(2013\)](#) and SBD of [Mandel et al. \(2015\)](#) place received rewards into queues for each arm, taking one whenever a base bandit algorithm suggests playing the arm. Throughout, we take UCB1 ([Auer et al., 2002](#)) as the base algorithm in QPM-D. [Joulani et al. \(2013\)](#) also present a direct modification of the UCB1 algorithm. All of these algorithms achieve the stated regret. None of them require *any* knowledge of the delay distributions, but they all rely heavily upon the non-anonymous nature of the observations.

While these results are encouraging, the assumption that the rewards are observed individually in a non-anonymous fashion is limiting for most practical applications with delays (e.g., recall the applications discussed earlier). How big is the price to be paid for receiving only aggregated anonymous feedback? Our main result is to prove that essentially there is no extra price to be paid provided that the value of the expected delay (or a bound on it) is available. In particular, this means that detailed knowledge of which action led to a particular delayed reward can be replaced by the much weaker requirement that the expected delay, or a bound on it, is known. Fig. 1 summarizes the relationship between the non-delayed, the delayed and the new problem

by showing the leading terms of the regret. In all cases, the dominant term is  $\sqrt{KT}$ . Hence, asymptotically, the delayed, aggregated anonymous feedback problem is no more difficult than the standard multi-armed bandit problem.

### 1.1. Our Techniques and Results

We now consider what sort of algorithm will be able to achieve the aforementioned results for the MABDAF problem. Since the player only observes delayed, aggregated anonymous rewards, the first problem we face is how to even estimate the mean reward of individual actions. Due to the delays and anonymity, it appears that to be able to estimate the mean reward of an action, the player wants to have played it consecutively for long stretches. Indeed, if the stretches are sufficiently long compared to the mean delay, the observations received during the stretch will mostly consist of rewards of the action played in that stretch. This naturally leads to considering algorithms that *switch actions rarely* and this is indeed the basis of our approach.

Several popular MAB algorithms are based on choosing the action with the largest upper confidence bound (UCB) in each round (e.g., [Auer et al., 2002](#); [Cappé et al., 2013](#)). UCB-style algorithms tend to switch arms frequently and will only play the optimal arm for long stretches if a unique optimal arm exists. Therefore, for MABDAF, we will consider alternative algorithms where arm-switching is more tightly controlled. The design of such algorithms goes back at least to the work of [Agrawal et al. \(1988\)](#) where the problem of bandits with switching costs was studied. The general idea of these rarely switching algorithms is to gradually eliminate suboptimal arms by playing arms in phases and comparing each arm’s upper confidence bound to the lower confidence bound of a leading arm at the end of each phase. Generally, this sort of rarely switching algorithm switches arms only  $O(\log T)$  times. We base our approach on one such algorithm, the so-called Improved UCB<sup>1</sup> algorithm of [Auer & Ortner \(2010\)](#).

Using a rarely switching algorithm alone will not be sufficient for MABDAF. The remaining problem, and where the bulk of our contribution lies, is to construct appropri-

<sup>1</sup>The adjective “Improved” indicates that the algorithm improves upon the regret bounds achieved by UCB1. The improvement replaces  $\log(T)/\Delta_j$  by  $\log(T\Delta_j^2)/\Delta_j$  in the regret bound.

ate confidence bounds and adjust the length of the periods of playing each arm to account for the delayed, aggregated anonymous feedback. In particular, in the confidence bounds attention must be paid to fine details: it turns out that unless the variance of the observations is dealt with, there is a blow-up by a multiplicative factor of  $K$ . We avoid this by an improved analysis involving Freedman’s inequality (Freedman, 1975). Further, to handle the dependencies between the number of plays of each arm and the past rewards, we combine Doob’s optimal skipping theorem (Doob, 1953) and Azuma-Hoeffding inequalities. Using a rarely switching algorithm for MABDAAF means we must also consider the dependencies between the elimination of arms in one phase and the corruption of observations in the next phase (ie. past plays can influence both whether an arm is still active and the corruption of its next plays). We deal with this through careful algorithmic design.

Using the above, we provide an algorithm that achieves worst case regret of  $O(\sqrt{KT \log K} + K\mathbb{E}[\tau] \log T)$  using only knowledge of the expected delay,  $\mathbb{E}[\tau]$ . We then show that this regret can be improved by using a more careful martingale argument that exploits the fact that our algorithm is designed to remove most of the dependence between the corruption of future observations and elimination of arms. Particularly, if the delays are bounded with known bound  $0 \leq d \leq \sqrt{T/K}$ , we can recover worst case regret of  $O(\sqrt{KT \log K} + K\mathbb{E}[\tau])$ , matching that of Joulani et al. (2013). If the delays are unbounded but have known variance  $\mathbb{V}(\tau)$ , we show that the problem independent regret can be reduced to  $O(\sqrt{KT \log K} + K\mathbb{E}[\tau] + K\mathbb{V}(\tau))$ .

## 1.2. Related Work

We have already discussed several of the most relevant works to our own. However, there has also been other work looking at different flavors of the bandit problem with delayed (non-anonymous) feedback. For example, Neu et al. (2010) and Cesa-Bianchi et al. (2016) consider non-stochastic bandits with fixed constant delays; Dudik et al. (2011) look at stochastic contextual bandits with a constant delay and Desautels et al. (2014) consider Gaussian Process bandits with a bounded stochastic delay. The general observation that delay causes an additive regret penalty in stochastic bandits and a multiplicative one in adversarial bandits is made in Joulani et al. (2013). The empirical performance of  $K$ -armed stochastic bandit algorithms in delayed settings was investigated in Chapelle & Li (2011). A further related problem is the ‘batched bandit’ problem studied by Perchet et al. (2016). Here the player must fix a set of time points at which to collect feedback on all plays leading up to that point. Vernade et al. (2017) consider delayed Bernoulli bandits where some observations could also be censored (e.g., no conversion is ever actually observed if the delay exceeds some threshold) but require

complete knowledge of the delay distribution. Crucially, here and in all the aforementioned works, the feedback is always assumed to take the form of arm-reward pairs and knowledge of the assignment of rewards to arms underpins the suggested algorithms, rendering them unsuitable for MABDAAF. To the best of our knowledge, ours is the first work to develop algorithms to deal with delayed, aggregated anonymous feedback in the bandit setting.

## 1.3. Organization

The remainder of this paper is organized as follows: In the next section (Section 2) we give the formal problem definition. We present our algorithm in Section 3. In Section 4, we discuss the performance of our algorithm under various delay assumptions; known expectation, bounded support with known bound and expectation, and known variance and expectation. This is followed by a numerical illustration of our results in Section 5. We conclude in Section 6.

## 2. Problem Definition

There are  $K > 1$  actions or arms in the set  $\mathcal{A}$ . Each action  $j \in \mathcal{A}$  is associated with a reward distribution  $\zeta_j$  and a delay distribution  $\delta_j$ . The reward distribution is supported in  $[0, 1]$  and the delay distribution is supported on  $\mathbb{N} \doteq \{0, 1, \dots\}$ . We denote by  $\mu_j$  the mean of  $\zeta_j$ ,  $\mu^* = \mu_{j^*} = \max_j \mu_j$  and define  $\Delta_j = \mu^* - \mu_j$  to be the *reward gap*, that is the expected loss of reward each time action  $j$  is chosen instead of an optimal action. Let  $(R_{l,j}, \tau_{l,j})_{l \in \mathbb{N}, j \in \mathcal{A}}$  be an infinite array of random variables defined on the probability space  $(\Omega, \Sigma, P)$  which are mutually independent. Further,  $R_{l,j}$  follows the distribution  $\zeta_j$  and  $\tau_{l,j}$  follows the distribution  $\delta_j$ . The meaning of these random variables is that if the player plays action  $j$  at time  $l$ , a payoff of  $R_{l,j}$  will be added to the aggregated feedback that the player receives at the end of the  $(l + \tau_{l,j})$ th play. Formally, if  $J_l \in \mathcal{A}$  denotes the action chosen by the player at time  $l = 1, 2, \dots$ , then the observation received at the end of the  $t$ th play is

$$X_t = \sum_{l=1}^t \sum_{j=1}^K R_{l,j} \times \mathbb{I}\{l + \tau_{l,j} = t, J_l = j\}.$$

For the remainder, we will consider i.i.d. delays across arms. We also assume discrete delay distributions, although most results hold for continuous delays by redefining the event  $\{\tau_{l,j} = t - l\}$  as  $\{t - l - 1 < \tau_{l,j} \leq t - l\}$  in  $X_t$ . In our analysis, we will sum over stochastic index sets. For a stochastic index set  $I$  and random variables  $\{Z_n\}_{n \in \mathbb{N}}$  we denote such sums as  $\sum_{t \in I} Z_t \doteq \sum_{t \in \mathbb{N}} \mathbb{I}\{t \in I\} \times Z_t$ .

**Regret definition** In most bandit problems, the regret is the cumulative loss due to not playing an optimal action.

In the case of delayed feedback, there are several possible ways to define the regret. One option is to consider only the loss of the rewards *received* before horizon  $T$  (as in [Vernade et al. \(2017\)](#)). However, we will not use this definition. Instead, as in [Joulani et al. \(2013\)](#), we consider the loss of all *generated* rewards and define the (pseudo-)regret by

$$\mathfrak{R}_T = \sum_{t=1}^T (\mu^* - \mu_{J_t}) = T\mu^* - \sum_{t=1}^T \mu_{J_t}.$$

This includes the rewards received after the horizon  $T$  and does not penalize large delays as long as an optimal action is taken. This definition is natural since, in practice, the player should eventually receive all outstanding reward.

[Lai & Robbins \(1985\)](#) showed that the regret of any algorithm for the standard MAB problem must satisfy,

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\mathfrak{R}_T]}{\log(T)} \geq \sum_{j: \Delta_j > 0} \frac{\Delta_j}{KL(\zeta_j, \zeta^*)}, \quad (1)$$

where  $KL(\zeta_j, \zeta^*)$  is the KL-divergence between the reward distributions of arm  $j$  and an optimal arm. Theorem 4 of [Vernade et al. \(2017\)](#) shows that the lower bound in (1) also holds for delayed feedback bandits with no censoring and their alternative definition of regret. We therefore suspect (1) should hold for MABDAAF. However, due to the specific problem structure, finding a lower bound for MABDAAF is non-trivial and remains an open problem.

**Assumptions on delay distribution** For our algorithm for MABDAAF, we need some assumptions on the delay distribution. We assume that the expected delay,  $\mathbb{E}[\tau]$ , is bounded and known. This quantity is used in the algorithm.

**Assumption 1** *The expected delay  $\mathbb{E}[\tau]$  is bounded and known to the algorithm.*

We then show that under some further mild assumptions on the delay, we can obtain better algorithms with even more efficient regret guarantees. We consider two settings: delay distributions with bounded support, and bounded variance.

**Assumption 2 (Bounded support)** *There exists some constant  $d > 0$  known to the algorithm such that the support of the delay distribution is bounded by  $d$ .*

**Assumption 3 (Bounded variance)** *The variance,  $\mathbb{V}(\tau)$ , of the delay is bounded and known to the algorithm.*

In fact the known expected value and known variance assumption can be replaced by a ‘known upper bound’ on the expected value and variance respectively. However, for simplicity, in the remaining, we use  $\mathbb{E}[\tau]$  and  $\mathbb{V}(\tau)$  directly. The next sections provide algorithms and regret analysis for different combinations of the above assumptions.

### 3. Our Algorithm

Our algorithm is a phase-based elimination algorithm based on the Improved UCB algorithm by [Auer & Ortner \(2010\)](#). The general structure is as follows. In each phase, each arm is played multiple times consecutively. At the end of the phase, the observations received are used to update mean estimates, and any arm with an estimated mean below the best estimated mean by a gap larger than a ‘separation gap tolerance’ is eliminated. This separation tolerance is decreased exponentially over phases, so that it is very small in later phases, eliminating all but the best arm(s) with high probability. An alternative formulation of the algorithm is that at the end of a phase, any arm with an upper confidence bound lower than the best lower confidence bound is eliminated. These confidence bounds are computed so that with high probability they are more (less) than the true mean, but within the separation gap tolerance. The phase lengths are then carefully chosen to ensure that the confidence bounds hold. Here we assume that the horizon  $T$  is known, but we expect that this can be relaxed as in [Auer & Ortner \(2010\)](#).

**Algorithm overview** Our algorithm, ODAAF, is given in Algorithm 1. It operates in phases  $m = 1, 2, \dots$ . Define  $\mathcal{A}_m$  to be the set of active arms in phase  $m$ . The algorithm takes parameter  $n_m$  which defines the number of samples of each active arm required by the end of phase  $m$ .

In Step 1 of phase  $m$  of the algorithm, each active arm  $j$  is played repeatedly for  $n_m - n_{m-1}$  steps. We record all timesteps where arm  $j$  was played in the first  $m$  phases (excluding bridge periods) in the set  $T_j(m)$ . The active arms are played in any arbitrary but fixed order. In Step 2, the  $n_m$  observations from timesteps in  $T_j(m)$  are averaged to obtain a new estimate  $\bar{X}_{m,j}$  of  $\mu_j$ . Arm  $j$  is eliminated if  $\bar{X}_{m,j}$  is further than  $\hat{\Delta}_m$  from  $\max_{j' \in \mathcal{A}_m} \bar{X}_{m,j'}$ .

A further nuance in the algorithm structure is the ‘bridge period’ (see Figure 2). The algorithm picks an active arm  $j \in \mathcal{A}_{m+1}$  to play in this bridge period for  $n_m - n_{m-1}$  steps. The observations received during the bridge period are discarded, and not used for computing confidence intervals. The significance of the bridge period is that it breaks the dependence between confidence intervals calculated in phase  $m$  and the delayed payoffs seeping into phase  $m+1$ . Without the bridge period this dependence would impair the validity of our confidence intervals. However, we suspect that, in practice, it may be possible to remove it.

**Choice of  $n_m$**  A key element of our algorithm design is the careful choice of  $n_m$ . Since  $n_m$  determines the number of times each active (possibly suboptimal) arm is played, it clearly has an impact on the regret. Furthermore,  $n_m$  needs to be chosen so that the confidence bounds on the estimation error hold with given probability. The main chal-

**Algorithm 1** Optimism for Delayed, Aggregated Anonymous Feedback (ODAAF)

**Input:** A set of arms,  $\mathcal{A}$ ; a horizon,  $T$ ; choice of  $n_m$  for each phase  $m = 1, 2, \dots$

**Initialization:** Set  $\tilde{\Delta}_1 = 1/2$  (tolerance), the set of active arms  $\mathcal{A}_1 = \mathcal{A}$ . Let  $T_i(1) = \emptyset, i \in \mathcal{A}, m = 1$  (phase index),  $t = 1$  (round index)

**while**  $t \leq T$  **do**

Step 1: Play arms.

**for**  $j \in \mathcal{A}_m$  **do**

Let  $T_j(m) = T_j(m-1)$

**while**  $|T_j(m)| \leq n_m$  **and**  $t \leq T$  **do**

Play arm  $j$ , receive  $X_t$ . Add  $t$  to  $T_j(m)$ . Increment  $t$  by 1.

**end while**

**end for**

Step 2: Eliminate sub-optimal arms.

For every arm in  $j \in \mathcal{A}_m$ , compute  $\bar{X}_{m,j}$  as the average of observations at time steps  $t \in T_j(m)$ . That is,

$$\bar{X}_{m,j} = \frac{1}{|T_j(m)|} \sum_{t \in T_j(m)} X_t.$$

Construct  $\mathcal{A}_{m+1}$  by eliminating actions  $j \in \mathcal{A}_m$  with

$$\bar{X}_{m,j} + \tilde{\Delta}_m < \max_{j' \in \mathcal{A}_m} \bar{X}_{m,j'}.$$

Step 3: Decrease Tolerance.

Set  $\tilde{\Delta}_{m+1} = \frac{\tilde{\Delta}_m}{2}$ .

Step 4: Bridge period.

Pick an arm  $j \in \mathcal{A}_{m+1}$  and play it  $\nu_m = n_m - n_{m-1}$  times while incrementing  $t \leq T$ . Discard all observations from this period. Do not add  $t$  to  $T_j(m)$ .

Increment phase index  $m$ .

**end while**

lenge is developing these confidence bounds from delayed, aggregated anonymous feedback. Handling this form of feedback involves a credit assignment problem of deciding which samples can be used for a given arm's mean estimation, since each sample is an aggregate of rewards from multiple previously played arms. This credit assignment problem would be hopeless in a passive learning setting without further information on how the samples were generated. Our algorithm utilizes the power of active learning to design the phases in such a way that the feedback can be effectively 'decensored' without losing too many samples.

A naive approach to defining the confidence bounds for delays bounded by a constant  $d \geq 0$  would be to observe that,

$$\left| \sum_{t \in T_j(m) \setminus T_j(m-1)} X_t - \sum_{t \in T_j(m) \setminus T_j(m-1)} R_{t,j} \right| \leq d,$$

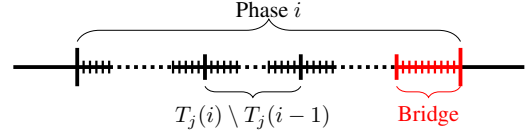


Figure 2: An example of phase  $i$  of our algorithm.

since all rewards are in  $[0, 1]$ . Then we could use Hoeffding's inequality to bound  $R_{t,j_t}$  (see Appendix F) and select

$$n_m = \frac{C_1 \log(T \tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} + \frac{C_2 m d}{\tilde{\Delta}_m}$$

for some constants  $C_1, C_2$ . This corresponds to worst case regret of  $O(\sqrt{KT \log K} + K \log(T)d)$ . For  $d \gg \mathbb{E}[\tau]$  and large  $T$ , this is significantly worse than that of Joulani et al. (2013). In Section 4, we show that, surprisingly, it is possible to recover the same rate of regret as Joulani et al. (2013), but this requires a significantly more nuanced argument to get tighter confidence bounds and smaller  $n_m$ . In the next section, we describe this improved choice of  $n_m$  for every phase  $m \in \mathbb{N}$  and its implications on the regret, for each of the three cases mentioned previously: (i) Known and bounded expected delay (Assumption 1), (ii) Bounded delay with known bound and expected value (Assumptions 1 and 2), (iii) Delay with known and bounded variance and expectation (Assumptions 1 and 3).

## 4. Regret Analysis

In this section, we specify the choice of parameters  $n_m$  and provide regret guarantees for Algorithm 1 for each of the three previously mentioned cases.

### 4.1. Known and Bounded Expected Delay

First, we consider the setting with the weakest assumption on delay distribution: we only assume that the expected delay,  $\mathbb{E}[\tau]$ , is bounded and known. No assumption on the support or variance of the delay distribution is made. The regret analysis for this setting will not use the bridge period, so Step 4 of the algorithm could be omitted in this case.

**Choice of  $n_m$**  Here, we use Algorithm 1 with

$$n_m = \frac{C_1 \log(T \tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} + \frac{C_2 m \mathbb{E}[\tau]}{\tilde{\Delta}_m} \quad (2)$$

for some large enough constants  $C_1, C_2$ . The exact value of  $n_m$  is given in Equation (14) in Appendix B.

**Estimation of error bounds** We bound the error between  $\bar{X}_{m,j}$  and  $\mu_j$  by  $\tilde{\Delta}_m/2$ . In order to do this we first bound the corruption of the observations received during timesteps  $T_j(m)$  due to delays.

Fix a phase  $m$  and arm  $j \in \mathcal{A}_m$ . Then the observations  $X_t$  in the period  $t \in T_j(m) \setminus T_j(m-1)$  are composed of two types of rewards: a subset of rewards from plays of arm  $j$  in this period, and delayed rewards from some of the plays before this period. The expected value of observations from this period would be  $(n_m - n_{m-1})\mu_j$  but for the rewards entering and leaving this period due to delay. Since the reward is bounded by 1, a simple observation is that expected discrepancy between the sum of observations in this period and the quantity  $(n_m - n_{m-1})\mu_j$  is bounded by the expected delay  $\mathbb{E}[\tau]$ ,

$$\mathbb{E} \left[ \sum_{t \in T_j(m) \setminus T_j(m-1)} (X_t - \mu_j) \right] \leq \mathbb{E}[\tau]. \quad (3)$$

Summing this over phases  $\ell = 1, \dots, m$  gives a bound

$$|\mathbb{E}[\bar{X}_{m,j}] - \mu_j| \leq \frac{m\mathbb{E}[\tau]}{|T_j(m)|} = \frac{m\mathbb{E}[\tau]}{n_m}. \quad (4)$$

Note that given the choice of  $n_m$  in (2), the above is smaller than  $\tilde{\Delta}_m/2$ , when large enough constants are used. Using this, along with concentration inequalities and the choice of  $n_m$  from (2), we can obtain the following high probability bound. A detailed proof is provided in Appendix B.1.

**Lemma 1** *Under Assumption 1 and the choice of  $n_m$  given by (2), the estimates  $\bar{X}_{m,j}$  constructed by Algorithm 1 satisfy the following: For every fixed arm  $j$  and phase  $m$ , with probability  $1 - \frac{3}{T\tilde{\Delta}_m^2}$ , either  $j \notin \mathcal{A}_m$ , or:*

$$|\bar{X}_{m,j} - \mu_j| \leq \tilde{\Delta}_m/2.$$

**Regret bounds** Using Lemma 1, we derive the following regret bounds in the current setting.

**Theorem 2** *Under Assumption 1, the expected regret of Algorithm 1 is upper bounded as*

$$\mathbb{E}[\mathfrak{R}_T] \leq \sum_{\substack{j=1 \\ j \neq j^*}}^K O \left( \frac{\log(T\tilde{\Delta}_j^2)}{\Delta_j} + \log(1/\Delta_j)\mathbb{E}[\tau] \right). \quad (5)$$

*Proof:* Given Lemma 1, the proof of Theorem 2 closely follows the analysis of the Improved UCB algorithm of Auer & Ortner (2010). Lemma 1 and the elimination condition in Algorithm 1 ensure that, with high probability, any suboptimal arm  $j$  will be eliminated by phase  $m_j = \log(1/\Delta_j)$ , thus incurring regret at most  $n_{m_j}\Delta_j$ . We then substitute in  $n_{m_j}$  from (2), and sum over all suboptimal arms. A detailed proof is in Appendix B.2. As in Auer & Ortner (2010), we avoid a union bound over all arms (which would result in an extra  $\log K$ ) by (i) reasoning about the regret of each arm individually, and (ii) bounding the regret resulting

from erroneously eliminating the optimal arm by carefully controlling the probability it is eliminated in each phase.  $\square$

Considering the worst-case values of  $\Delta_j$  (roughly  $\sqrt{K/T}$ ), we obtain the following problem independent bound.

**Corollary 3** *For any problem instance satisfying Assumption 1, the expected regret of Algorithm 1 satisfies*

$$\mathbb{E}[\mathfrak{R}_T] \leq O(\sqrt{KT \log(K)} + K\mathbb{E}[\tau] \log(T)).$$

## 4.2. Delay with Bounded Support

If the delay is bounded by some constant  $d \geq 0$  and a single arm is played repeatedly for long enough, we can restrict the number of arms corrupting the observation  $X_t$  at a given time  $t$ . In fact, if each arm  $j$  is played consecutively for more than  $d$  rounds, then at any time  $t \in T_j(m)$ , the observation  $X_t$  will be composed of the rewards from at most two arms: the current arm  $j$ , and previous arm  $j'$ . Further, from the elimination condition, with high probability, arm  $j'$  will have been eliminated if it is clearly suboptimal. We can then recursively use the confidence bounds for arms  $j$  and  $j'$  from the previous phase to bound  $|\mu_j - \mu_{j'}|$ . Below, we formalize this intuition to obtain a tighter bound on  $|\bar{X}_{m,j} - \mu_j|$  for every arm  $j$  and phase  $m$ , when each active arm is played a specified number of times per phase.

**Choice of  $n_m$**  Here, we define,

$$n_m = \frac{C_1 \log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} + \frac{C_2 \mathbb{E}[\tau]}{\tilde{\Delta}_m} + \min \left\{ md, \frac{C_3 \log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} + \frac{C_4 m \mathbb{E}[\tau]}{\tilde{\Delta}_m} \right\} \quad (6)$$

for some large enough constants  $C_1, C_2, C_3, C_4$  (see Appendix C, Equation (18) for the exact values). This choice of  $n_m$  means that for large  $d$ , we essentially revert back to the choice of  $n_m$  from (2) for the unbounded case, and we gain nothing by using the bound on the delay. However, if  $d$  is not large, the choice of  $n_m$  in (6) is smaller than (2) since the second term now scales with  $\mathbb{E}[\tau]$  rather than  $m\mathbb{E}[\tau]$ .

**Estimation of error bounds** In this setting, by the elimination condition and bounded delays, the expectation of each reward entering  $T_j(m)$  will be within  $\tilde{\Delta}_{m-1}$  of  $\mu_j$ , with high probability. Then, using knowledge of the upper bound of the support of  $\tau$ , we can obtain a tighter bound and get an error bound similar to Lemma 1 with the smaller value of  $n_m$  in (6). We prove the following proposition. Since  $\tilde{\Delta}_m = 2^{-m}$ , this is considerably tighter than (3).

**Proposition 4** *Assume  $n_i - n_{i-1} \geq d$  for phases  $i = 1, \dots, m$ . Define  $\mathcal{E}_{m-1}$  as the event that all arms  $j \in \mathcal{A}_m$  satisfy error bounds  $|\bar{X}_{m-1,j} - \mu_j| \leq \tilde{\Delta}_{m-1}/2$ . Then, for*

every arm  $j \in \mathcal{A}_m$ ,

$$\mathbb{E} \left[ \sum_{t \in T_j(m) \setminus T_j(m-1)} (X_t - \mu_j) \middle| \mathcal{E}_{m-1} \right] \leq \tilde{\Delta}_{m-1} \mathbb{E}[\tau].$$

*Proof:* (Sketch). Consider a fixed arm  $j \in \mathcal{A}_m$ . The expected value of the sum of observations  $X_t$  for  $t \in T_j(m) \setminus T_j(m-1)$  would be  $(n_m - n_{m-1})\mu_j$  were it not for some rewards entering and leaving this period due to the delays. Because of the i.i.d. assumption on the delay, in expectation, the number of rewards leaving the period is roughly the same as the number of rewards entering this period, i.e.,  $\mathbb{E}[\tau]$ . (Conditioning on  $\mathcal{E}_{m-1}$  does not effect this due to the bridge period). Since  $n_m - n_{m-1} \geq d$ , the reward coming into the period  $T_j(m) \setminus T_j(m-1)$  can only be from the previous arm  $j'$ . All rewards leaving the period are from arm  $j$ . Therefore the expected difference between rewards entering and leaving the period is  $(\mu_j - \mu_{j'})\mathbb{E}[\tau]$ . Then, if  $\mu_j$  is close to  $\mu_{j'}$ , the total reward leaving the period is compensated by total reward entering. Due to the bridge period, even when  $j$  is the first arm played in phase  $m$ ,  $j' \in \mathcal{A}_m$ , so it was not eliminated in phase  $m-1$ . By the elimination condition in Algorithm 1, if the error bounds  $|\bar{X}_{m-1,j} - \mu_j| \leq \tilde{\Delta}_{m-1}/2$  are satisfied for all arms in  $\mathcal{A}_m$ , then  $|\mu_j - \mu_{j'}| \leq \tilde{\Delta}_{m-1}$ . This gives the result.  $\square$

Repeatedly using Proposition 4 we get,

$$\sum_{i=1}^m \mathbb{E} \left[ \sum_{t \in T_j(i) \setminus T_j(i-1)} (X_t - \mu_j) \middle| \mathcal{E}_{i-1} \right] \leq 2\mathbb{E}[\tau]$$

since  $\sum_{i=1}^m \tilde{\Delta}_{i-1} = \sum_{i=0}^{m-1} 2^{-i} \leq 2$ . Then, observe that  $\mathbb{P}(\mathcal{E}_i^C)$  is small. This bound is an improvement of a factor of  $m$  compared to (4). For the regret analysis, we derive a high probability version of the above result. Using this, and the choice of  $n_m \geq \Omega\left(\frac{\log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} + \frac{\mathbb{E}[\tau]}{\tilde{\Delta}_m}\right)$  from (6), for large enough constants, we derive the following lemma. A detailed proof is given in Appendix C.1.

**Lemma 5** *Under Assumptions 1 of known expected delay and 2 of bounded delays, and choice of  $n_m$  given in (6), the estimates  $\bar{X}_{m,j}$  obtained by Algorithm 1 satisfy the following: For any arm  $j$  and phase  $m$ , with probability at least  $1 - \frac{12}{T\tilde{\Delta}_m^2}$ , either  $j \notin \mathcal{A}_m$  or*

$$|\bar{X}_{m,j} - \mu_j| \leq \tilde{\Delta}_m/2.$$

**Regret bounds** We now give regret bounds for this case.

**Theorem 6** *Under Assumption 1 and bounded delay Assumption 2, the expected regret of Algorithm 1 satisfies*

$$\mathbb{E}[\mathfrak{R}_T] \leq \sum_{j=1: j \neq j^*}^K O\left(\frac{\log(T\Delta_j^2)}{\Delta_j} + \mathbb{E}[\tau]\right)$$

$$+ \min\left\{d, \frac{\log(T\Delta_j^2)}{\Delta_j} + \log\left(\frac{1}{\Delta_j}\right)\mathbb{E}[\tau]\right\}.$$

*Proof:* (Sketch). Given Lemma 5, the proof is similar to that of Theorem 2. The full proof is in Appendix C.2.  $\square$

Then, if  $d \leq \sqrt{\frac{T \log K}{K}} + \mathbb{E}[\tau]$ , we get the following problem independent regret bound which matches that of Joulani et al. (2013).

**Corollary 7** *For any problem instance satisfying Assumptions 1 and 2 with  $d \leq \sqrt{\frac{T \log K}{K}} + \mathbb{E}[\tau]$ , the expected regret of Algorithm 1 satisfies*

$$\mathbb{E}[\mathfrak{R}_T] \leq O(\sqrt{KT \log(K)} + K\mathbb{E}[\tau]).$$

### 4.3. Delay with Bounded Variance

If the delay is unbounded but well behaved in the sense that we know (a bound on) the variance, then we can obtain similar regret bounds to the bounded delay case. Intuitively, delays from the previous phase will only corrupt observations in the current phase if their delays exceed the length of the bridge period. We control this by using the bound on the variance to bound the tails of the delay distributions.

**Choice of  $n_m$**  Let  $\mathbb{V}(\tau)$  be the known variance (or bound on the variance) of the delay, as in Assumption 3. Then, we use Algorithm 1 with the following value of  $n_m$ ,

$$n_m = C_1 \frac{\log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} + C_2 \frac{\mathbb{E}[\tau] + \mathbb{V}(\tau)}{\tilde{\Delta}_m} \quad (7)$$

for some large enough constants  $C_1, C_2$ . The exact value of  $n_m$  is given in Appendix D, Equation (25).

**Regret bounds** We get the following instance specific and problem independent regret bound in this case.

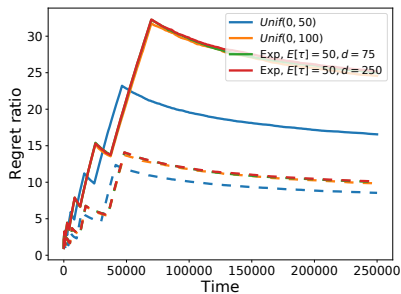
**Theorem 8** *Under Assumption 1 and Assumption 3 of known (bound on) the expectation and variance of the delay, and choice of  $n_m$  from (7), the expected regret of Algorithm 1 can be upper bounded by,*

$$\mathbb{E}[\mathfrak{R}_T] \leq \sum_{j=1: \mu_j \neq \mu^*}^K O\left(\frac{\log(T\Delta_j^2)}{\Delta_j} + \mathbb{E}[\tau] + \mathbb{V}(\tau)\right).$$

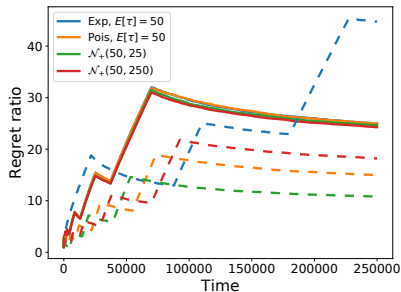
*Proof:* (Sketch). See Appendix D.2. We use Chebychev's inequality to get a result similar to Lemma 5 and then use a similar argument to the bounded delay case.  $\square$

**Corollary 9** *For any problem instance satisfying Assumptions 1 and 3, the expected regret of Algorithm 1 satisfies*

$$\mathbb{E}[\mathfrak{R}_T] \leq O(\sqrt{KT \log(K)} + K\mathbb{E}[\tau] + K\mathbb{V}(\tau)).$$



(a) Bounded delays. Ratios of regret of ODAAF (solid lines) and ODAAF-B (dotted lines) to that of QPM-D.



(b) Unbounded delays. Ratios of regret of ODAAF (solid lines) and ODAAF-V (dotted lines) to that of QPM-D.

Figure 3: The ratios of regret of variants of our algorithm to that of QPM-D for different delay distributions.

**Remark** If  $\mathbb{E}[\tau] \geq 1$ , then the delay penalty can be reduced to  $O(K\mathbb{E}[\tau] + K\mathbb{V}(\tau)/\mathbb{E}[\tau])$  (see Appendix D).

Thus, it is sufficient to know a bound on variance to obtain regret bounds similar to those in bounded delay case. Note that this approach is not possible just using knowledge of the expected delay since we cannot guarantee that the reward entering phase  $i$  is from an arm active in phase  $i - 1$ .

## 5. Experimental Results

We compared the performance of our algorithm (under different assumptions) to QPM-D (Joulani et al., 2013) in various experimental settings. In these experiments, our aim was to investigate the effect of the delay on the performance of the algorithms. In order to focus on this, we used a simple setup of two arms with Bernoulli rewards and  $\mu = (0.5, 0.6)$ . In every experiment, we ran each algorithm to horizon  $T = 250000$  and used UCB1 (Auer et al., 2002) as the base algorithm in QPM-D. The regret was averaged over 200 replications. For ease of reading, we define ODAAF to be our algorithm using only knowledge of the expected delay, with  $n_m$  defined as in (2) and run without a bridge period, and ODAAF-B and ODAAF-V to be the versions of Algorithm 1 that use a bridge period and information on the bounded support and the finite variance of the delay to define  $n_m$  as in (6) and (7) respectively.

We tested the algorithms with different delay distributions. In the first case, we considered bounded delay distributions whereas in the second case, the delays were unbounded. In Fig. 3a, we plotted the ratios of the regret of ODAAF and ODAAF-B (with knowledge of  $d$ , the delay bound) to the regret of QPM-D. We see that in all cases the ratios converge to a constant. This shows that the regret of our algorithm is essentially of the same order as that of QPM-D. Our algorithm predetermines the number of times to play each active arm per phase (the randomness appears in whether an arm is active), so the jumps in the regret are it changing arm. This occurs at the same points in all replications.

Fig. 3b shows a similar story for unbounded delays with mean  $\mathbb{E}[\tau] = 50$  (where  $\mathcal{N}_+$  denotes the the half normal distribution). The ratios of the regret of ODAAF and ODAAF-V (with knowledge of the delay variance) to the regret of QPM-D again converge to constants. Note that in this case, these constants, and the location of the jumps, vary with the delay distribution and  $\mathbb{V}(\tau)$ . When the variance of the delay is small, it can be seen that using the variance information leads to improved performance. However, for exponential delays where  $\mathbb{V}(\tau) = \mathbb{E}[\tau]^2$ , the large variance causes  $n_m$  to be large and so the suboptimal arm is played more, increasing the regret. In this case ODAAF-V had only just eliminated the suboptimal arm at time  $T$ .

It can also be illustrated experimentally that the regret of our algorithms and that of QPM-D all increase linearly in  $\mathbb{E}[\tau]$ . This is shown in Appendix E. We also provide an experimental comparison to Vernade et al. (2017) in Appendix E.

## 6. Conclusion

We have studied an extension of the multi-armed bandit problem to bandits with delayed, aggregated anonymous feedback. Here, a sum of observations is received after some stochastic delay and we do not learn which arms contributed to each observation. In this more difficult setting, we have proven that, surprisingly, it is possible to develop an algorithm that performs comparably to those for the simpler delayed feedback bandits problem, where the assignment of rewards to plays is known. Particularly, using only knowledge of the expected delay, our algorithm matches the worst case regret of Joulani et al. (2013) up to a logarithmic factor. This logarithmic factors can be removed using an improved analysis and slightly more information about the delay; if the delay is bounded, we achieve the same worst case regret as Joulani et al. (2013), and for unbounded delays with known finite variance, we have an extra additive  $\mathbb{V}(\tau)$  term. We supported these claims experimentally. Note that while our algorithm matches the order of regret of QPM-D, the constants are worse. Hence, it is an open problem to find algorithms with better constants.



## Acknowledgments

CPB would like to thank the EPSRC funded EP/L015692/1 STOR-i centre for doctoral training and Sparx. We would like to thank the reviewers for their helpful comments.

## References

- Agrawal, R., Hedge, M., and Teneketzis, D. Asymptotically efficient adaptive allocation rules for the multi-armed bandit problem with switching cost. *IEEE Transactions on Automatic Control*, 33(10):899–906, 1988.
- Auer, P. and Ortner, R. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.
- Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Journal of Machine Learning Research*, 47(2-3):235–256, 2002.
- Bubeck, S. and Cesa-Bianchi, N. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Foundations and Trends in Machine Learning. 2012.
- Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.
- Cesa-Bianchi, N., Gentile, C., Mansour, Y., and Minora, A. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pp. 605–622, 2016.
- Chapelle, O. and Li, L. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, pp. 2249–2257, 2011.
- Desautels, T., Krause, A., and Burdick, J. Parallelizing exploration-exploitation tradeoffs in Gaussian process bandit optimization. *Journal of Machine Learning Research*, 15(1):3873–3923, 2014.
- Doob, J. L. *Stochastic processes*. John Wiley & Sons, 1953.
- Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. Efficient optimal learning for contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, pp. 169–178, 2011.
- Freedman, D. A. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.
- Joulani, P., György, A., and Szepesvári, C. Online learning under delayed feedback. In *International Conference on Machine Learning*, pp. 1453–1461, 2013.
- Lai, T. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–22, 1985.
- Mandel, T., Liu, Y.-E., Brunskill, E., and Popovic, Z. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *AAAI*, pp. 2849–2856, 2015.
- Neu, G., Antos, A., György, A., and Szepesvári, C. Online Markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, pp. 1804–1812, 2010.
- Perchet, V., Rigollet, P., Chassang, S., and Snowberg, E. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681, 2016.
- Szita, I. and Szepesvári, C. Agnostic KWIK learning and efficient approximate reinforcement learning. In *Conference on Learning Theory*, pp. 739–772, July 2011.
- Thompson, W. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3–4):285–294, 1933.
- Vernade, C., Cappé, O., and Perchet, V. Stochastic bandit models for delayed conversions. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

## Appendix

### A. Preliminaries

#### A.1. Table of Notation

For ease of reading, we define here key notation that will be used in this Appendix.

$T$	: The horizon.
$\Delta_j$	: The gap between the mean of the optimal arm and the mean of arm $j$ , $\Delta_j = \mu^* - \mu_j$ .
$\tilde{\Delta}_m$	: The approximation to $\Delta_j$ at round $m$ of the ODAAF algorithm, $\tilde{\Delta}_m = \frac{1}{2^m}$ .
$n_m$	: The number of samples of an active arm $j$ ODAAF needs by the end of round $m$ .
$\nu_m$	: The number of times each arm is played in phase $m$ , $\nu_m = n_m - n_{m-1}$ .
$d$	: The bound on the delay in the case of bounded delay.
$m_j$	: The first round of the ODAAF algorithm where $\tilde{\Delta}_m < \Delta_j/2$ .
$M_j$	: The random variable representing the round arm $j$ is eliminated in.
$T_j(m)$	: The set of all time point where arm $j$ is played up to (and including) round $m$ .
$X_t$	: The reward received at time $t$ (from any possible past plays of the algorithm).
$R_{t,j}$	: The reward generated by playing arm $j$ at time $t$ .
$\tau_{t,j}$	: The delay associated with playing arm $j$ at time $t$ .
$\mathbb{E}[\tau]$	: The expected delay (assuming i.i.d. delays).
$\mathbb{V}(\tau)$	: The variance of the delay (assuming i.i.d. delays).
$\bar{X}_{m,j}$	: The estimated reward of arm $j$ in phase $m$ . See Algorithm 1 for the definition.
$S_m$	: The start point of the $m$ th phase. See Appendix A.2 for more details.
$U_m$	: The end point of the $m$ th phase. See Appendix A.2 for more details.
$S_{m,j}$	: The start point of phase $m$ of playing arm $j$ . See Appendix A.2 for more details.
$U_{m,j}$	: The end point of phase $m$ of playing arm $j$ . See Appendix A.2 for more details.
$\mathcal{A}_m$	: The set of active arms in round $m$ of the ODAAF algorithm.
$A_{i,t}, B_{i,t}, C_{i,t}$	: The contribution of the reward generated at time $t$ in certain intervals relating to phase $i$ to the corruption. See (11) for the exact definitions.
$\mathcal{G}_t$	: The smallest $\sigma$ -algebra containing all information up to time $t$ , see (8) for a definition.

#### A.2. Beginning and End of Phases

We formalize here some notation that will be used throughout the analysis to denote the start and end points of each phase. Define the random variables  $S_i$  and  $U_i$  for each phase  $i = 1, \dots, m$  to be the start and end points of the phase. Then let  $S_{i,j}, U_{i,j}$  denote the start and end points of playing arm  $j$  in phase  $i$ . See Figure 4 for details. By convention, let  $S_{i,j} = U_{i,j} = \infty$  if arm  $j$  is not active in phase  $i$ ,  $S_i = U_i = \infty$  if the algorithm never reaches phase  $i$  and let  $S_{0,j} = U_{0,j} = S_0 = U_0 = 0$  for all  $j$ . It is important to point out that  $n_m$  are deterministic so at the end of any phase  $m - 1$ , once we have eliminated sub-optimal arms, we also know which arms are in  $\mathcal{A}_m$  and consequently the start and end points of phase  $m$ . Furthermore, since we play arms in a given order, we also know the specific rounds when we start and finish playing each active arm in phase  $m$ . Hence, at any time step  $t$  in phase  $m$ ,  $S_m, U_m, S_{m+1}$  and  $U_{m,j}, S_{m,j}$  for all active arms  $j \in \mathcal{A}_m$  will be known. More formally, define the filtration  $\{\mathcal{G}_t\}_{t=0}^\infty$  where

$$\mathcal{G}_t = \sigma(X_1, \dots, X_t, \tau_{1,J_1}, \dots, \tau_{t,J_t}, R_{1,J_1}, \dots, R_{t,J_t}, J_1, \dots, J_t) \quad (8)$$

and  $\mathcal{G}_0 = \{\emptyset, \Omega\}$ . This means the joint events like  $\{S_i \leq t\} \cap \{S_{i,j} = s'\} \in \mathcal{G}_t$  for all  $s' \in \mathbb{N}, j \in \mathcal{A}$ .

#### A.3. Useful Results

For our analysis, we will need Freedman's version of Bernstein's inequality for the right-tail of martingales with bounded increments:

**Theorem 10 (Freedman's version of Bernstein's inequality; Theorem 1.6 of Freedman (1975))** *Let  $\{Y_k\}_{k=0}^\infty$  be a real-valued martingale with respect to the filtration  $\{\mathcal{F}_k\}_{k=0}^\infty$  with increments  $\{Z_k\}_{k=1}^\infty$ :  $\mathbb{E}[Z_k | \mathcal{F}_{k-1}] = 0$  and  $Z_k = Y_k - Y_{k-1}$ , for  $k = 1, 2, \dots$ . Assume that the difference sequence is uniformly bounded on the right:  $Z_k \leq b$  almost surely for  $k = 1, 2, \dots$ . Define the predictable variation process  $W_k = \sum_{j=1}^k \mathbb{E}[Z_j^2 | \mathcal{F}_{j-1}]$  for  $k = 1, 2, \dots$ . Then, for all  $t \geq 0$ ,*

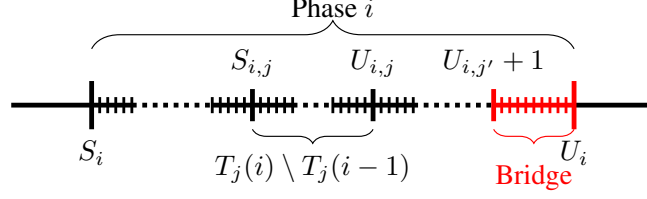


Figure 4: An example of phase  $i$  of our algorithm. Here  $j'$  is the last active arm played in phase  $i$ .

$\sigma^2 > 0$ ,

$$\mathbb{P}(\exists k \geq 0 : Y_k \geq t \text{ and } W_k \leq \sigma^2) \leq \exp\left\{-\frac{t^2/2}{\sigma^2 + bt/3}\right\}.$$

This result implies that if for some deterministic constant,  $\sigma^2$ ,  $W_k \leq \sigma^2$  holds almost surely, then  $\mathbb{P}(Y_k \geq t) \leq \exp\left\{-\frac{t^2/2}{\sigma^2 + bt/3}\right\}$  holds for any  $t \geq 0$ .

We will also make use of the following technical lemma which combines the Hoeffding-Azuma inequality and Doob's optional skipping theorem (Theorem 2.3 in Chapter VII of [Doob \(1953\)](#)):

**Lemma 11** Fix the positive integers  $m, n$  and let  $a, c \in \mathbb{R}$ . Let  $\mathcal{F} = \{\mathcal{F}_t\}_{t=0}^n$  be a filtration,  $(\epsilon_t, Z_t)_{t=1,2,\dots,n}$  be a sequence of  $\{0, 1\} \times \mathbb{R}$ -valued random variables such that for  $t \in \{1, 2, \dots, n\}$ ,  $\epsilon_t$  is  $\mathcal{F}_{t-1}$ -measurable,  $Z_t$  is  $\mathcal{F}_t$ -measurable,  $\mathbb{E}[Z_t | \mathcal{F}_{t-1}] = 0$  and  $Z_t \in [a, a + c]$ . Further, assume that  $\sum_{s=1}^n \epsilon_s \leq m$  with probability one. Then, for any  $\lambda > 0$ ,

$$\mathbb{P}\left(\sum_{t=1}^n \epsilon_t Z_t \geq \lambda\right) \leq \exp\left\{-\frac{2\lambda^2}{c^2 m}\right\}. \quad (9)$$

*Proof:* This lemma appeared in a slightly more general form (where  $n = \infty$  is allowed) as Lemma A.1 in the paper by [Szita & Szepesvári \(2011\)](#) so we refer the reader to the proof there.  $\square$

## B. Results for Known and Bounded Expected Delay

### B.1. High Probability Bounds

**Lemma 1** Under Assumption 1 and the choice of  $n_m$  given by (2), the estimates  $\bar{X}_{m,j}$  constructed by Algorithm 1 satisfy the following: For every fixed arm  $j$  and phase  $m$ , with probability  $1 - \frac{3}{T\tilde{\Delta}_m^2}$ , either  $j \notin \mathcal{A}_m$ , or:

$$\bar{X}_{m,j} - \mu_j \leq \tilde{\Delta}_m/2.$$

*Proof:* Let

$$w_m = \frac{4 \log(T\tilde{\Delta}_m^2)}{3n_m} + \sqrt{\frac{2 \log(T\tilde{\Delta}_m^2)}{n_m} + \frac{3m\mathbb{E}[\tau]}{n_m}}. \quad (10)$$

We first show that with probability greater than  $1 - \frac{3}{T\tilde{\Delta}_m^2}$ ,  $j \notin \mathcal{A}_m$  or  $\frac{1}{n_m} \sum_{t \in T_j(m)} (X_t - \mu_j) \leq w_m$ .

For arm  $j$  and phase  $m$ , assume  $j \in \mathcal{A}_m$ . For notational simplicity we will use in the following  $\mathbb{I}_i\{H\} := \mathbb{I}\{H \cap \{j \in \mathcal{A}_i\}\} \leq \mathbb{I}\{H\}$  for any event  $H$ . If  $j \in \mathcal{A}_m$  for a particular experiment  $\omega$  then  $\mathbb{I}_i(H)(\omega) = \mathbb{I}(H)(\omega)$ . Then for any phase  $i \leq m$  and time  $t$ , define,

$$A_{i,t} = R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t \geq S_i\}, \quad B_{i,t} = R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\}, \quad C_{i,t} = R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t > U_{i,j}\}, \quad (11)$$

and note that since  $S_{i,j} = U_{i,j} = \infty$  if arm  $j$  is not active in phase  $i$ , we have the equalities  $\mathbb{I}_i\{\tau_{t,J_t} + t \geq S_{i,j}\} = \mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\}$  and  $\mathbb{I}_i\{\tau_{t,J_t} + t > U_{i,j}\} = \mathbb{I}\{\tau_{t,J_t} + t > U_{i,j}\}$ . Define the filtration  $\{\mathcal{G}_s\}_{s=0}^\infty$  by  $\mathcal{G}_0 = \{\Omega, \emptyset\}$  and

$$\mathcal{G}_t = \sigma(X_1, \dots, X_t, J_1, \dots, J_t, \tau_{1,J_1}, \dots, \tau_{t,J_t}, R_{1,J_1}, \dots, R_{t,J_t}). \quad (12)$$

Then, we use the decomposition,

$$\begin{aligned}
 \sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} (X_t - \mu_j) &\leq \sum_{i=1}^m \left( \sum_{t=S_{i-1,j}}^{S_{i,j}-1} R_{t,J_t} \mathbb{I}_i \{ \tau_{t,J_t} + t \geq S_{i,j} \} + \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j) - \sum_{t=S_{i,j}}^{U_{i,j}} R_{t,J_t} \mathbb{I}_i \{ \tau_{t,J_t} + t > U_{i,j} \} \right) \\
 &\leq \sum_{i=1}^m \left( \sum_{t=S_{i-1,j}}^{S_i-1} R_{t,J_t} \mathbb{I} \{ \tau_{t,J_t} + t \geq S_i \} + \sum_{t=S_i}^{S_{i,j}-1} R_{t,J_t} \mathbb{I} \{ \tau_{t,J_t} + t \geq S_{i,j} \} \right. \\
 &\quad \left. + \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j) - \sum_{t=S_{i,j}}^{U_{i,j}} R_{t,J_t} \mathbb{I} \{ \tau_{t,J_t} + t > U_{i,j} \} \right) \\
 &= \sum_{i=1}^m \left( \sum_{t=S_{i-1,j}}^{S_i-1} A_{i,t} + \sum_{t=S_i}^{S_{i,j}-1} B_{i,t} + \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j) - \sum_{t=S_{i,j}}^{U_{i,j}} C_{i,t} \right) \\
 &= \sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j) + \sum_{t=1}^{S_{m,j}} Q_t - \sum_{t=1}^{U_{m,j}} P_t \\
 &= \underbrace{\sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j)}_{\text{Term I.}} + \underbrace{\sum_{t=1}^{S_{m,j}} (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}])}_{\text{Term II.}} + \underbrace{\sum_{t=1}^{U_{m,j}} (\mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t)}_{\text{Term III.}} \\
 &\quad + \underbrace{\left( \sum_{t=1}^{S_{m,j}} \mathbb{E}[Q_t | \mathcal{G}_{t-1}] - \sum_{t=1}^{U_{m,j}} \mathbb{E}[P_t | \mathcal{G}_{t-1}] \right)}_{\text{Term IV.}},
 \end{aligned} \tag{13}$$

where,

$$\begin{aligned}
 Q_t &= \sum_{i=1}^m (A_{i,t} \mathbb{I} \{ S_{i-1,j} \leq t \leq S_i - 1 \} + B_{i,t} \mathbb{I} \{ S_i \leq t \leq S_{i,j} - 1 \}) \\
 P_t &= \sum_{i=1}^m C_{i,t} \mathbb{I} \{ S_{i,j} \leq t \leq U_{i,j} \}.
 \end{aligned}$$

Recall that the filtration  $\{\mathcal{G}_s\}_{s=0}^\infty$  is defined by  $\mathcal{G}_0 = \{\Omega, \emptyset\}$ ,  $\mathcal{G}_t = \sigma(X_1, \dots, X_t, J_1, \dots, J_t, \tau_{1,J_1}, \dots, \tau_{t,J_t}, R_{1,J_1}, \dots, R_{t,J_t})$  and we have defined  $S_{i,j} = \infty$  if arm  $j$  is eliminated before phase  $i$  and  $S_i = \infty$  if the algorithm stops before reaching phase  $i$ .

**Outline of proof** We will bound each term of the above decomposition in (13) in turn, however first we need to prove several intermediary results. For term II., we will use Freedman's inequality so we first need Lemma 12 to show that  $Z_t = Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]$  is a martingale difference and Lemma 13 to bound the variance of the sum of the  $Z_t$ 's. Similarly, for term III., in Lemma 14, we show that  $Z'_t = \mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t$  is a martingale difference and bound its variance in Lemma 15. In Lemma 16, we consider term IV. and bound the conditional expectations of  $A_{i,t}, B_{i,t}, C_{i,t}$ . Finally, in Lemma 17, we bound term I. using Lemma 11. We then combine the bounds on all terms together to conclude the proof.

**Lemma 12** *Let  $Y_s = \sum_{t=1}^s (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}])$  for all  $s \geq 1$ ,  $Y_0 = 0$ . Then  $\{Y_s\}_{s=0}^\infty$  is a martingale with respect to the filtration  $\{\mathcal{G}_s\}_{s=0}^\infty$  with increments  $Z_s = Y_s - Y_{s-1} = Q_s - \mathbb{E}[Q_s | \mathcal{G}_{s-1}]$  satisfying  $\mathbb{E}[Z_s | \mathcal{G}_{s-1}] = 0$ ,  $Z_s \leq 1$  for all  $s \geq 1$ .*

*Proof:* To show  $\{Y_s\}_{s=0}^\infty$  is a martingale with respect to  $\{\mathcal{G}_s\}_{s=0}^\infty$ , we need to show that  $Y_s$  is  $\mathcal{G}_s$  measurable for all  $s$  and  $\mathbb{E}[Y_s | \mathcal{G}_{s-1}] = Y_{s-1}$ .

**Measurability:** First note that by definition of  $\mathcal{G}_s$ ,  $\tau_{t,J_t}, R_{t,J_t}$  are all  $\mathcal{G}_s$ -measurable for  $t \leq s$ . Then, for each  $i$ , either  $t$  is in a phase later than  $i$  so  $S_{i-1,j}$  and  $S_i$  are  $\mathcal{G}_t$ -measurable, or  $S_{i-1,j}$  and  $S_i$  are not  $\mathcal{G}_t$ -measurable, but  $\mathbb{I}\{t \geq S_{i,j}\} = 0$  so  $\mathbb{I}\{t \geq S_{i,j}\}$  is  $\mathcal{G}_t$ -measurable. In the first case, since  $S_{i-1,j}$  and  $S_i$  are  $\mathcal{G}_t$ -measurable  $A_{i,t} \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - \nu_i\}$  is  $\mathcal{G}_t$ -measurable. In the second case,  $A_{i,t} \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - 1\} = A_{i,t} \mathbb{I}\{S_{i-1,j} \leq t\} \mathbb{I}\{t \leq S_i - 1\} = 0$  so it is also

$\mathcal{G}_t$ -measurable. Similarly, if  $t$  is after  $S_i$ ,  $S_i$  and  $S_{i,j}$  will be  $\mathcal{G}$ -measurable or  $\mathbb{I}\{S_i \leq t \leq S_{i,j} - 1\} = 0$ . In both cases,  $B_{i,t}\mathbb{I}\{S_i \leq t \leq S_{i,j} - 1\}$  is  $\mathcal{G}_t$ -measurable. Hence,  $Q_t$  is  $\mathcal{G}_t$ -measurable, and also  $Q_t$  is  $\mathcal{G}_s$  measurable for any  $s \geq t$ . It then follows that  $Y_s$  is  $\mathcal{G}_s$ -measurable for all  $s$ .

Expectation: Since  $Q_t$  is  $\mathcal{G}_s$  measurable for all  $t \leq s$ ,

$$\begin{aligned} \mathbb{E}[Y_s | \mathcal{G}_{s-1}] &= \mathbb{E}\left[\sum_{t=1}^s (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]) | \mathcal{G}_{s-1}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{s-1} (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]) | \mathcal{G}_{s-1}\right] + \mathbb{E}[(Q_s - \mathbb{E}[Q_s | \mathcal{G}_{s-1}]) | \mathcal{G}_{s-1}] \\ &= \sum_{t=1}^{s-1} (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]) + \mathbb{E}[Q_s | \mathcal{G}_{s-1}] - \mathbb{E}[Q_s | \mathcal{G}_{s-1}] \\ &= \sum_{t=1}^{s-1} (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]) = Y_{s-1} \end{aligned}$$

Hence,  $\{Y_s\}_{s=0}^\infty$  is a martingale with respect to the filtration  $\{\mathcal{G}_s\}_{s=0}^\infty$ .

Increments: For any  $s = 1, \dots$ , we have that

$$Z_s = Y_s - Y_{s-1} = \sum_{t=1}^s (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]) - \sum_{t=1}^{s-1} (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]) = Q_s - \mathbb{E}[Q_s | \mathcal{G}_{s-1}].$$

Then,

$$\mathbb{E}[Z_s | \mathcal{G}_{s-1}] = \mathbb{E}[Q_s - \mathbb{E}[Q_s | \mathcal{G}_{s-1}] | \mathcal{G}_{s-1}] = \mathbb{E}[Q_s | \mathcal{G}_{s-1}] - \mathbb{E}[Q_s | \mathcal{G}_{s-1}] = 0.$$

Lastly, since for any  $t$ , there is only one  $i$  where one of  $\mathbb{I}\{S_{i-1,j} \leq t \leq S_i - 1\} = 1$  or  $\mathbb{I}\{S_i \leq t \leq S_{i,j} - 1\} = 1$  (and they cannot both be one), and since  $R_{t,J_t} \in [0, 1]$ ,  $A_{i,t}, B_{i,t} \leq 1$ , so it follows that  $Z_s = Q_s - \mathbb{E}[Q_s | \mathcal{G}_{s-1}] \leq 1$  for all  $s$ .  $\square$

**Lemma 13** For any  $t$ , let  $Z_t = Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]$ , then, for any  $s < S_{m,j}$ ,

$$\sum_{t=1}^s \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] \leq 2m\mathbb{E}[\tau].$$

*Proof:* First note that

$$\begin{aligned} \sum_{t=1}^s \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] &= \sum_{t=1}^s \mathbb{V}(Q_t | \mathcal{G}_{t-1}) \leq \sum_{t=1}^s \mathbb{E}[Q_t^2 | \mathcal{G}_{t-1}] \\ &= \sum_{t=1}^s \mathbb{E}\left[\left(\sum_{i=1}^m (A_{i,t}\mathbb{I}\{S_{i-1,j} \leq t \leq S_i - 1\} + B_{i,t}\mathbb{I}\{S_i \leq t \leq S_{i,j} - 1\})\right)^2 \middle| \mathcal{G}_{t-1}\right]. \end{aligned}$$

Then, given  $\mathcal{G}_{t-1}$ , all indicator terms  $\mathbb{I}\{S_{i-1,j} \leq t \leq S_i - 1\}$  and  $\mathbb{I}\{S_i \leq t \leq S_{i,j} - 1\}$  for all  $i = 1, \dots, m$  are measurable and only one can be non zero. Hence, all interaction terms in the expansion of the quadratic are 0 and so we are left with

$$\begin{aligned} \sum_{t=1}^s \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] &\leq \sum_{t=1}^s \mathbb{E}\left[\left(\sum_{i=1}^m (A_{i,t}\mathbb{I}\{S_{i-1,j} \leq t \leq S_i - 1\} + B_{i,t}\mathbb{I}\{S_i \leq t \leq S_{i,j} - 1\})\right)^2 \middle| \mathcal{G}_{t-1}\right] \\ &= \sum_{t=1}^s \mathbb{E}\left[\sum_{i=1}^m (A_{i,t}^2 \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - 1\}^2 + B_{i,t}^2 \mathbb{I}\{S_i \leq t \leq S_{i,j} - 1\}^2) \middle| \mathcal{G}_{t-1}\right] \\ &= \sum_{i=1}^m \sum_{t=1}^s \mathbb{E}[A_{i,t}^2 \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - 1\} | \mathcal{G}_{t-1}] + \sum_{i=1}^m \sum_{t=1}^s \mathbb{E}[B_{i,t}^2 \mathbb{I}\{S_i \leq t \leq S_{i,j} - 1\} | \mathcal{G}_{t-1}] \end{aligned}$$

$$\leq \sum_{i=1}^m \sum_{t=S_{i-1,j}}^{S_i-1} \mathbb{E}[A_{i,t}^2 | \mathcal{G}_{t-1}] + \sum_{i=1}^m \sum_{t=S_i}^{S_{i,j}-1} \mathbb{E}[B_{i,t}^2 | \mathcal{G}_{t-1}].$$

Then, for any  $i \geq 1$ ,

$$\begin{aligned} \sum_{t=S_{i-1,j}}^{S_i-1} \mathbb{E}[A_{i,t}^2 | \mathcal{G}_{t-1}] &= \sum_{t=S_{i-1,j}}^{S_i-1} \mathbb{E}[R_{t,J_t}^2 \mathbb{I}\{\tau_{t,J_t} + t \geq S_i\} | \mathcal{G}_{t-1}] \\ &\leq \sum_{t=S_{i-1,j}}^{S_i-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_i\} | \mathcal{G}_{t-1}] \\ &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i-1,j} = s, S_i = s'\} \sum_{t=s}^{s'-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_i\} | \mathcal{G}_{t-1}] \\ &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s}^{s'-1} \mathbb{E}[\mathbb{I}\{S_{i-1,j} = s, S_i = s', \tau_{t,J_t} + t \geq S_i\} | \mathcal{G}_{t-1}] \\ &\hspace{15em} \text{(Since } \{t \geq S_{i-1,j}, S_i = s'\} \in \mathcal{G}_{t-1}\text{)} \\ &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s}^{s'-1} \mathbb{E}[\mathbb{I}\{S_{i-1,j} = s, S_i = s', \tau_{t,J_t} + t \geq s'\} | \mathcal{G}_{t-1}] \\ &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i-1,j} = s, S_i = s'\} \sum_{t=s}^{s'-1} \mathbb{P}(\tau_{t,J_t} + t \geq s') \\ &\hspace{15em} \text{(Since } \{t \geq S_{i-1,j}, S_i = s'\} \in \mathcal{G}_{t-1}\text{)} \\ &\leq \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i-1,j} = s, S_i = s'\} \sum_{l=0}^{\infty} \mathbb{P}(\tau > l) \\ &\leq \mathbb{E}[\tau]. \end{aligned}$$

Likewise, for any  $i \geq 1$ ,

$$\begin{aligned} \sum_{t=S_i}^{S_{i,j}-1} \mathbb{E}[B_{i,t}^2 | \mathcal{G}_{t-1}] &= \sum_{t=S_i}^{S_{i,j}-1} \mathbb{E}[R_{t,J_t}^2 \mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\} | \mathcal{G}_{t-1}] \\ &\leq \sum_{t=S_i}^{S_{i,j}-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\} | \mathcal{G}_{t-1}] \\ &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_i = s, S_{i,j} = s'\} \sum_{t=s}^{s'-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\} | \mathcal{G}_{t-1}] \\ &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s}^{s'-1} \mathbb{E}[\mathbb{I}\{S_i = s, S_{i,j} = s', \tau_{t,J_t} + t \geq S_{i,j}\} | \mathcal{G}_{t-1}] \\ &\hspace{15em} \text{(Since } \{t \geq S_i, S_{i,j} = s'\} \in \mathcal{G}_{t-1}\text{)} \\ &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s}^{s'-1} \mathbb{E}[\mathbb{I}\{S_i = s, S_{i,j} = s', \tau_{t,J_t} + t \geq s'\} | \mathcal{G}_{t-1}] \\ &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_i = s, S_{i,j} = s'\} \sum_{t=s}^{s'-1} \mathbb{P}(\tau_{t,J_t} + t \geq s') \\ &\hspace{15em} \text{(Since } \{t \geq S_i, S_{i,j} = s'\} \in \mathcal{G}_{t-1}\text{)} \\ &\leq \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_i = s, S_{i,j} = s'\} \sum_{l=0}^{\infty} \mathbb{P}(\tau \geq l) \end{aligned}$$

$$\leq \mathbb{E}[\tau].$$

Hence, combining both terms and summing over the phases  $m$  gives the result.  $\square$

**Lemma 14** *Let  $Y'_s = \sum_{t=1}^s (\mathbb{E}[P_s | \mathcal{G}_{s-1}] - P_s)$  for all  $s \geq 1$ ,  $Y'_0 = 0$ . Then  $\{Y'_s\}_{s=0}^\infty$  is a martingale with respect to the filtration  $\{\mathcal{G}_s\}_{s=0}^\infty$  with increments  $Z'_s = Y'_s - Y'_{s-1} = \mathbb{E}[P_s | \mathcal{G}_{s-1}] - P_s$  satisfying  $\mathbb{E}[Z'_s | \mathcal{G}_{s-1}] = 0$ ,  $Z'_s \leq 1$  for all  $s \geq 1$ .*

*Proof:* The proof is similar to that of Lemma 12. To show  $\{Y'_s\}_{s=0}^\infty$  is a martingale with respect to  $\{\mathcal{G}_s\}_{s=0}^\infty$ , we need to show that  $Y'_s$  is  $\mathcal{G}_s$  measurable for all  $s$  and  $\mathbb{E}[Y'_s | \mathcal{G}_{s-1}] = Y'_{s-1}$ .

Measurability: As before, by definition of  $\mathcal{G}_s$ ,  $\tau_{t,J_t}, R_{t,J_t}$  are all  $\mathcal{G}_s$ -measurable for  $t \leq s$ . Also, we can reduce measurability again to measurability of  $\mathbb{I}\{\tau_{s,J_s} + s \geq U_{i,j}, S_{i,j} \leq s \leq U_{i,j}\}$ . But,  $\{U_{i,j} = s'\} \cap \{S_{i,j} \leq s\} \in \mathcal{G}_s$  for all  $s' \in \mathbb{N}$  and  $Y'_s$  is adapted to  $\mathcal{G}_s$ .

Increments: For any  $s \geq 1$ , we have that

$$Z'_s = Y'_s - Y'_{s-1} = \sum_{t=1}^s (\mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t) - \sum_{t=1}^{s-1} (\mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t) = \mathbb{E}[P_s | \mathcal{G}_{s-1}] - P_s.$$

Then,

$$\mathbb{E}[Z'_s | \mathcal{G}_{s-1}] = \mathbb{E}[\mathbb{E}[P_s | \mathcal{G}_{s-1}] - P_s | \mathcal{G}_{s-1}] = \mathbb{E}[P_s | \mathcal{G}_{s-1}] - \mathbb{E}[P_s | \mathcal{G}_{s-1}] = 0.$$

Lastly, since for any  $t$  and  $\omega \in \Omega$ , there is at most one  $i$  for which  $\mathbb{I}\{S_{i,j} \leq t \leq U_{i,j}\} = 1$ , and by definition of  $R_{t,J_t}$ ,  $C_{i,t} \leq 1$ , so it follows that  $Z'_s = \mathbb{E}[P_s | \mathcal{G}_{s-1}] - P_s \leq 1$  for all  $s$ .  $\square$

**Lemma 15** *For any  $t$ , let  $Z'_t = \mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t$ , then*

$$\sum_{t=1}^{U_{m,j}} \mathbb{E}[Z_t'^2 | \mathcal{G}_{t-1}] \leq m \mathbb{E}[\tau].$$

*Proof:* The proof is similar to that of Lemma 13. First note that

$$\begin{aligned} \sum_{t=1}^{U_{m,j}} \mathbb{E}[Z_t'^2 | \mathcal{G}_{t-1}] &= \sum_{t=1}^{U_{m,j}} \mathbb{V}(P_t | \mathcal{G}_{t-1}) \leq \sum_{t=1}^{U_{m,j}} \mathbb{E}[P_t^2 | \mathcal{G}_{t-1}] \\ &= \sum_{t=1}^{U_{m,j}} \mathbb{E} \left[ \left( \sum_{i=1}^m (C_{i,t} \mathbb{I}\{S_{i,j} \leq t \leq U_{i,j}\}) \right)^2 \middle| \mathcal{G}_{t-1} \right]. \end{aligned}$$

Then, given  $\mathcal{G}_{t-1}$ , all indicator terms  $\mathbb{I}\{S_{i,j} \leq t \leq U_{i,j}\}$  for  $i = 1, \dots, m$  are measurable and at most one can be non zero. Hence, all interaction terms are 0 and so we are left with

$$\begin{aligned} \sum_{t=1}^{U_{m,j}} \mathbb{E}[Z_t'^2 | \mathcal{G}_{t-1}] &\leq \sum_{t=1}^{U_{m,j}} \mathbb{E} \left[ \left( \sum_{i=1}^m (C_{i,t} \mathbb{I}\{S_{i,j} \leq t \leq U_{i,j}\}) \right)^2 \middle| \mathcal{G}_{t-1} \right] \\ &= \sum_{i=1}^m \sum_{t=1}^{U_{m,j}} \mathbb{E}[C_{i,t}^2 \mathbb{I}\{S_{i,j} \leq t \leq U_{i,j}\} | \mathcal{G}_{t-1}] \\ &\leq \sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} \mathbb{E}[C_{i,t}^2 | \mathcal{G}_{t-1}] \quad (\text{since the indicator is } \mathcal{G}_{t-1}\text{-measurable}) \\ &= \sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} \mathbb{E}[R_{t,J_t}^2 \mathbb{I}\{\tau_{t,J_t} + t > U_{i,j}\} | \mathcal{G}_{t-1}] \end{aligned}$$

$$\begin{aligned}
 &\leq \sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t > U_{i,j}\} | \mathcal{G}_{t-1}] \\
 &= \sum_{i=1}^m \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i,j} = s, U_{i,j} = s'\} \sum_{t=s}^{s'} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t > U_{i,j}\} | \mathcal{G}_{t-1}] \\
 &= \sum_{i=1}^m \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s}^{s'} \mathbb{E}[\mathbb{I}\{S_{i,j} = s, U_{i,j} = s', \tau_{t,J_t} + t > U_{i,j}\} | \mathcal{G}_{t-1}] \\
 &= \sum_{i=1}^m \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s}^{s'} \mathbb{E}[\mathbb{I}\{S_{i,j} = s, U_{i,j} = s', \tau_{t,J_t} + t > s'\} | \mathcal{G}_{t-1}] \\
 &= \sum_{i=1}^m \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i,j} = s, U_{i,j} = s'\} \sum_{t=s}^{s'} \mathbb{P}(\tau_{t,J_t} + t > s') \\
 &\leq \sum_{i=1}^m \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i,j} = s, U_{i,j} = s'\} \sum_{l=0}^{\infty} \mathbb{P}(\tau > l) \\
 &\leq \sum_{i=1}^m \mathbb{E}[\tau] = m\mathbb{E}[\tau].
 \end{aligned}$$

□

**Lemma 16** For  $A_{i,t}, B_{i,t}$  and  $C_{i,t}$  defined as in (11), let  $\nu_i = n_i - n_{i-1}$  be the number of times each arm is played in phase  $i$  and  $j'_i$  be the arm played directly before arm  $j$  in phase  $i$ . Then, it holds that, for any arm  $j$  and phase  $i \geq 1$ ,

$$\begin{aligned}
 (i) \quad &\sum_{t=S_{i-1,j}}^{S_i-1} \mathbb{E}[A_{i,t} | \mathcal{G}_{t-1}] \leq \mathbb{E}[\tau] \\
 (ii) \quad &\sum_{t=S_i}^{S_{i,j}-1} \mathbb{E}[B_{i,t} | \mathcal{G}_{t-1}] \leq \mathbb{E}[\tau] + \mu_{j'_i} \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l) \\
 (iii) \quad &\sum_{t=S_{i,j}}^{U_{i,j}} \mathbb{E}[C_{i,t} | \mathcal{G}_{t-1}] = \mu_j \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l)
 \end{aligned}$$

*Proof:* We prove each statement individually. Several of the proofs are similar to those appearing in Lemmas 13 and 15.

**Statement (i):**

$$\begin{aligned}
 \sum_{t=S_{i-1,j}}^{S_i-1} \mathbb{E}[A_{i,t} | \mathcal{G}_{t-1}] &\leq \sum_{t=S_{i-1,j}}^{S_i-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_i\} | \mathcal{G}_{t-1}] \\
 &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i-1,j} = s, S_i = s'\} \sum_{t=s}^{s'-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_i\} | \mathcal{G}_{t-1}] \\
 &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s}^{s'-1} \mathbb{E}[\mathbb{I}\{S_{i-1,j} = s, S_i = s', \tau_{t,J_t} + t \geq S_i\} | \mathcal{G}_{t-1}] \\
 &\hspace{15em} (\text{Since } \{t \geq S_{i-1,j}, S_i = s'\} \in \mathcal{G}_{t-1}) \\
 &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s}^{s'-1} \mathbb{E}[\mathbb{I}\{S_{i-1,j} = s, S_i = s', \tau_{t,J_t} + t \geq s'\} | \mathcal{G}_{t-1}]
 \end{aligned}$$



$$\begin{aligned}
 &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i-1,j} = s, S_i = s'\} \sum_{t=s}^{s'-1} \mathbb{P}(\tau_{t,J_t} + t \geq s') \\
 &\hspace{15em} \text{(Since } \{t \geq S_{i-1,j}, S_i = s'\} \in \mathcal{G}_{t-1}\text{)} \\
 &\leq \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i-1,j} = s, S_i = s'\} \sum_{l=0}^{\infty} \mathbb{P}(\tau > l) \\
 &= \sum_{l=0}^{\infty} \mathbb{P}(\tau > l) = \mathbb{E}[\tau].
 \end{aligned}$$

**Statement (iii):**

$$\begin{aligned}
 \sum_{t=S_{i,j}}^{U_{i,j}} \mathbb{E}[C_{i,t} | \mathcal{G}_{t-1}] &= \sum_{t=S_{i,j}}^{U_{i,j}} \mathbb{E}[R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t > U_{i,j}\} | \mathcal{G}_{t-1}] \\
 &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i,j} = s, U_{i,j} = s'\} \sum_{t=s}^{s'} \mathbb{E}[R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t > U_{i,j}\} | \mathcal{G}_{t-1}] \\
 &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s}^{s'} \mathbb{E}[R_{t,J_t} \mathbb{I}\{S_{i,j} = s, U_{i,j} = s', \tau_{t,J_t} + t > U_{i,j}\} | \mathcal{G}_{t-1}] \\
 &\hspace{15em} \text{(Since } \{S_{i,j} = s, U_{i,j} = s'\} \in \mathcal{G}_{t-1} \text{ for } s \leq t\text{)} \\
 &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s}^{s'} \mathbb{E}[R_{t,J_t} \mathbb{I}\{S_{i,j} = s, U_{i,j} = s', \tau_{t,J_t} + t > s'\} | \mathcal{G}_{t-1}] \\
 &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i,j} = s, U_{i,j} = s'\} \sum_{t=s}^{s'} \mu_j \mathbb{P}(\tau_{t,J_t} + t > s') \\
 &\hspace{15em} \text{(Since } \{S_{i,j} = s, U_{i,j} = s'\} \in \mathcal{G}_{t-1} \text{ and given } \mathcal{G}_{t-1}, R_{t,J_t} \text{ and } \tau_{t,J_t} \text{ are independent)} \\
 &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i,j} = s, U_{i,j} = s'\} \mu_j \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l) \\
 &= \mu_j \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l)
 \end{aligned}$$

**Statement (ii):** For statement (ii), we have that for  $(i, j) \neq (1, 1)$ ,

$$\sum_{t=S_i}^{S_{i,j}-1} \mathbb{E}[B_{i,t} | \mathcal{G}_{t-1}] = \sum_{t=S_i}^{S_{i,j}-\nu_{i-1}-2} \mathbb{E}[B_{i,t} | \mathcal{G}_{t-1}] + \sum_{t=S_{i,j}-\nu_{i-1}-1}^{S_{i,j}-1} \mathbb{E}[B_{i,t} | \mathcal{G}_{t-1}].$$

Then,  $S_{i,j}$  is  $\mathcal{G}_{t-1}$  measurable for  $t \geq S_i$ , so we can use the same technique as for statement (i) to bound the first term. For the second term, since we will only be playing arm  $j'_i$  for  $S_{i,j} - \nu_{i-1} - 1, \dots, S_{i,j} - 1$ , we can use the same technique as for statement (iii). Hence,

$$\sum_{t=S_i}^{S_{i,j}-1} \mathbb{E}[B_{i,t} | \mathcal{G}_{t-1}] \leq \sum_{l=\nu_{i-1}+1}^{\infty} \mathbb{P}(\tau > l) + \mu_{j'_i} \sum_{l=0}^{\nu_{i-1}} \mathbb{P}(\tau > l) \leq \mathbb{E}[\tau] + \mu_{j'_i} \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l).$$

Note that, for  $(i, j) = (1, 1)$ , the amount seeping in will be 0, so using  $\nu_0 = 0, \mu'_{1,1} = 0$ , the result trivially holds. Hence the result holds for all  $i, j \geq 1$ .  $\square$

**Lemma 17** For any arm  $j \in \{1, \dots, K\}$  and phase  $m$ , it holds that for any  $\lambda > 0$ ,

$$\mathbb{P}\left(\sum_{t \in T_j(m)} (R_{t,j} - \mu_j) \geq \lambda\right) \leq \exp\left\{-\frac{2\lambda^2}{n_m}\right\}.$$

*Proof:* The result follows from Lemma 11. When applying this lemma, we use  $n = T$ ,  $m = n_m$ , for  $t = 0, 1, \dots, T$  set  $\mathcal{F}_t = \sigma(X_1, \dots, X_t, R_{1,j}, \dots, R_{t,j})$  and for  $t = 1, 2, \dots, T$  define  $Z_t = R_{t,j} - \mu_j$  and  $\epsilon_t = \mathbb{I}\{J_t = j, t \leq U_{m,j}\}$ . Note that  $T_j(m) = \{t \in \{1, \dots, T\} : \epsilon_t = 1\}$  and hence  $\sum_{t \in T_j(m)} (R_{t,j} - \mu_j) = \sum_{t=1}^T \epsilon_t (R_{t,j} - \mu_j)$ . Further,  $\sum_{t=1}^T \epsilon_t = |T_j(m)| \leq n_m$  with probability one.

Fix  $1 \leq t \leq T$ . We now argue that  $\epsilon_t$  is  $\mathcal{F}_{t-1}$ -measurable. First, notice that by the definition of ODAF, the index  $M$  of the phase that  $t$  belongs to can be calculated based on the observations  $X_1, \dots, X_{t-1}$  up to time  $t-1$ . Since  $t \leq U_{m,j}$  is equivalent to whether for this phase index  $M$ , the inequality  $M \leq m$  holds, it follows that  $\{t \leq U_{m,j}\}$  is  $\mathcal{F}_{t-1}$ -measurable. The same holds for  $\{J_t = j\}$  for the same reason. Hence, it follows that  $\epsilon_t$  is indeed  $\mathcal{F}_{t-1}$ -measurable.

Now,  $Z_t$  is  $\mathcal{F}_t$ -measurable as  $R_{t,j}$  is clearly  $\mathcal{F}_t$ -measurable. Furthermore, by our assumptions on  $(R_{t,j})_{t,j}$  and  $(X_t)_t$ ,  $\mathbb{E}[R_{t,j} | \mathcal{F}_{t-1}] = \mu_j$  also holds, implying that  $Z_t$  also satisfies the conditions of the lemma with  $a = -\mu_j$  and  $c = 1$ . Thus, the result follows by applying Lemma 11.  $\square$

We now bound each term of the decomposition in (13) in turn.

**Bounding Term I:** For Term I, we use Lemma 17 to get that with probability greater than  $1 - \frac{1}{T\tilde{\Delta}_m^2}$ ,

$$\sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j) \leq \sqrt{\frac{n_m \log(T\tilde{\Delta}_m^2)}{2}}.$$

**Bounding Term II:** For Term II, we will use Freedman's inequality (Theorem 10). From Lemma 12,  $\{Y_s\}_{s=0}^\infty$  with  $Y_s = \sum_{t=1}^s (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}])$  is a martingale with respect to  $\{\mathcal{G}_s\}_{s=0}^\infty$  with increments  $\{Z_s\}_{s=0}^\infty$  satisfying  $\mathbb{E}[Z_s | \mathcal{G}_{s-1}] = 0$  and  $Z_s \leq 1$  for all  $s$ . Further, by Lemma 13,  $\sum_{t=1}^s \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] \leq 2m\mathbb{E}[\tau] \leq \frac{6m \times 2^m \mathbb{E}[\tau]}{12} \leq n_m/12$  with probability 1. Hence we can apply Freedman's inequality to get that with probability greater than  $1 - \frac{1}{T\tilde{\Delta}_m^2}$ ,

$$\sum_{t=1}^{S_{m,j}} (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]) \leq \frac{2}{3} \log(T\tilde{\Delta}_m^2) + \sqrt{\frac{1}{12} n_m \log(T\tilde{\Delta}_m^2)}.$$

**Bounding Term III:** For Term III, we again use Freedman's inequality (Theorem 10) but using Lemma 14 to show that  $\{Y'_s\}_{s=0}^\infty$  with  $Y'_s = \sum_{t=1}^s (\mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t)$  is a martingale with respect to  $\{\mathcal{G}_s\}_{s=0}^\infty$  with increments  $\{Z'_s\}_{s=0}^\infty$  satisfying  $\mathbb{E}[Z'_s | \mathcal{G}_{s-1}] = 0$  and  $Z'_s \leq 1$  for all  $s$ . Further, by Lemma 15,  $\sum_{t=1}^s \mathbb{E}[Z'_t | \mathcal{G}_{t-1}] \leq m\mathbb{E}[\tau] \leq n_m/12$  with probability 1. Hence, with probability greater than  $1 - \frac{1}{T\tilde{\Delta}_m^2}$ ,

$$\sum_{t=1}^{U_{m,j}} (\mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t) \leq \frac{2}{3} \log(T\tilde{\Delta}_m) + \sqrt{\frac{1}{12} n_m \log(T\tilde{\Delta}_m^2)}.$$

**Bounding Term IV:** We bound term IV. using Lemma 16,

$$\begin{aligned} & \sum_{t=1}^{S_{m,j}} \mathbb{E}[Q_t | \mathcal{G}_{t-1}] - \sum_{t=1}^{U_{m,j}} \mathbb{E}[P_t | \mathcal{G}_{t-1}] \\ &= \sum_{t=1}^{S_{m,j}} \mathbb{E} \left[ \sum_{i=1}^m (A_{i,t} \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - 1\} + B_{i,t} \mathbb{I}\{S_i \leq t \leq S_{i,j} - 1\}) \middle| \mathcal{G}_{t-1} \right] \\ & \quad - \sum_{t=1}^{U_{m,j}} \mathbb{E} \left[ \sum_{i=1}^m C_{i,t} \mathbb{I}\{S_{i,j} \leq t \leq U_{i,j}\} \middle| \mathcal{G}_{t-1} \right] \\ &= \sum_{i=1}^m \sum_{t=1}^{S_{m,j}} \mathbb{E}[A_{i,t} \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - 1\} | \mathcal{G}_{t-1}] + \sum_{i=1}^m \sum_{t=1}^{S_{m,j}} \mathbb{E}[B_{i,t} \mathbb{I}\{S_i \leq t \leq S_{i,j} - 1\} | \mathcal{G}_{t-1}] \\ & \quad - \sum_{i=1}^m \sum_{t=1}^{U_{m,j}} \mathbb{E}[C_{i,t} \mathbb{I}\{S_{i,j} \leq t \leq U_{i,j}\} | \mathcal{G}_{t-1}] \end{aligned}$$

$$\begin{aligned}
 &= \sum_{i=1}^m \left( \sum_{t=S_{i-1},j}^{S_i-1} \mathbb{E}[A_{i,t}|\mathcal{G}_{t-1}] + \sum_{t=S_i}^{S_{i,j}-1} \mathbb{E}[B_{i,t}|\mathcal{G}_{t-1}] - \sum_{t=S_{i,j}}^{U_{i,j}} \mathbb{E}[C_{i,t}|\mathcal{G}_{t-1}] \right) \\
 &\leq \sum_{i=1}^m \left( 2\mathbb{E}[\tau] + \mu_{j_i} \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l) - \mu_j \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l) \right) \leq 3m\mathbb{E}[\tau].
 \end{aligned}$$

since  $R_{t,j} \in [0, 1]$ .

**Combining all terms:** To get the final high probability bound, we sum the bounds for each term I.-IV.. Then, with probability greater than  $1 - \frac{3}{T\tilde{\Delta}_m^2}$ , either  $j \notin \mathcal{A}_m$  or arm  $j$  is played  $n_m$  times by the end of phase  $m$  and

$$\begin{aligned}
 \frac{1}{n_m} \sum_{t \in T_j(m)} (X_t - \mu_j) &\leq \frac{4 \log(T\tilde{\Delta}_m^2)}{3n_m} + \left( \frac{2}{\sqrt{12}} + \frac{1}{\sqrt{2}} \right) \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{n_m}} + \frac{3m\mathbb{E}[\tau]}{n_m} \\
 &\leq \frac{4 \log(T\tilde{\Delta}_m^2)}{3n_m} + \sqrt{\frac{2 \log(T\tilde{\Delta}_m^2)}{n_m}} + \frac{3m\mathbb{E}[\tau]}{n_m} = w_m.
 \end{aligned}$$

**Defining  $n_m$ :** Setting

$$n_m = \left\lceil \frac{1}{\tilde{\Delta}_m^2} \left( \sqrt{2 \log(T\tilde{\Delta}_m^2)} + \sqrt{2 \log(T\tilde{\Delta}_m^2) + \frac{8}{3} \tilde{\Delta}_m \log(T\tilde{\Delta}_m^2) + 6\tilde{\Delta}_m m \mathbb{E}[\tau]} \right)^2 \right\rceil. \quad (14)$$

ensures that  $w_m \leq \frac{\tilde{\Delta}_m}{2}$  which concludes the proof.  $\square$

## B.2. Regret Bounds

Here we prove the regret bound in Theorem 2 under Assumption 1 and the choice of  $n_m$  given by (14). Under Assumption 1, the bridge period is not necessary so the results here hold for the version of Algorithm 1 with the bridge period omitted. Note that if we were to include the bridge period, we would be playing each arm at most  $2n_m$  times by the end of phase  $m$  so our regret would simply increase by a factor of 2.

**Theorem 2** *Under Assumption 1, the expected regret of Algorithm 1 is upper bounded as*

$$\mathbb{E}[\mathfrak{R}_T] \leq \sum_{\substack{j=1 \\ j \neq j^*}}^K O\left( \frac{\log(T\Delta_j^2)}{\Delta_j} + \log(1/\Delta_j)\mathbb{E}[\tau] \right). \quad (5)$$

*Proof:* Our proof is a restructuring of the proof of (Auer & Ortner, 2010). For any arm  $j$ , define  $M_j$  to be the random variable representing the phase when arm  $j$  is eliminated in. We set  $M_j = \infty$  if the arm did not get eliminated before time step  $T$ . Note that if  $M_j$  is finite,  $j \in \mathcal{A}_{M_j}$  (this also means that  $\mathcal{A}_{M_j}$  is well-defined) and if  $\mathcal{A}_{M_j+1}$  is also defined ( $M_j$  is not the last phase) then  $j \notin \mathcal{A}_{M_j+1}$ . We also let  $m_j$  denote the phase arm  $j$  should be eliminated in, that is  $m_j = \min\{m \geq 1 : \tilde{\Delta}_m < \frac{\Delta_j}{2}\}$ . From the definition of  $\tilde{\Delta}_m$  in our algorithm, we get the relations

$$2^{m_j} = \frac{1}{\tilde{\Delta}_{m_j}} \leq \frac{4}{\Delta_j} < \frac{1}{\tilde{\Delta}_{m_j+1}} \quad \text{and} \quad \frac{\Delta_j}{4} \leq \tilde{\Delta}_{m_j} \leq \frac{\Delta_j}{2}. \quad (15)$$

Define  $N_j = \sum_{t=1}^T \mathbb{I}\{J_t = j\}$  be the number of times arm  $j$  is used and let  $\mathfrak{R}_T^{(j)} = N_j \Delta_j$  be the ‘‘pseudo’’-regret contribution from each arm  $1 \leq j \leq K$  so that  $\mathbb{E}[\mathfrak{R}_T] = \mathbb{E}\left[\sum_{j=1}^K \mathfrak{R}_T^{(j)}\right]$ . Let  $M^*$  be the round when the optimal arm  $j^*$  is eliminated. Hence,

$$\mathbb{E}[\mathfrak{R}_T] = \mathbb{E}\left[\sum_{j=1}^K \mathfrak{R}_T^{(j)}\right] = \underbrace{\mathbb{E}\left[\sum_{j=1}^K \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* \geq m_j\}\right]}_{\text{Term I.}} + \underbrace{\mathbb{E}\left[\sum_{j=1}^K \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* < m_j\}\right]}_{\text{Term II.}}.$$

We will bound the regret in each of these cases in turn. To do so, we need the following results which consider the probabilities of confidence bounds failing and arms being eliminated in the incorrect rounds.

**Lemma 18** For any suboptimal arm  $j$ ,

$$\mathbb{P}(M_j > m_j \text{ and } M^* \geq m_j) \leq \frac{6}{T\tilde{\Delta}_{m_j}^2}.$$

*Proof:* Define

$$E = \{\bar{X}_{m_j,j} \leq \mu_j + w_{m_j}\} \quad \text{and} \quad H = \{\bar{X}_{m_j,j^*} > \mu^* - w_{m_j}\}.$$

If both  $E$  and  $F$  occur, it follows that,

$$\begin{aligned} \bar{X}_{m_j,j} &\leq \mu_j + w_{m_j} \\ &= \mu_j^* - \Delta_j + w_{m_j} && \text{(since } \Delta_j = \mu_{j^*} - \mu_j) \\ &\leq \bar{X}_{m_j,j^*} + w_{m_j} - \Delta_j + w_{m_j} \\ &< \bar{X}_{m_j,j^*} - 2\tilde{\Delta}_{m_j} + 2w_{m_j} && \text{(by (15))} \\ &\leq \bar{X}_{m_j,j^*} - \tilde{\Delta}_{m_j} && \text{(since } n_m \text{ is such that } w_m \leq \tilde{\Delta}_m/2) \end{aligned}$$

and arm  $j$  would be eliminated. Hence, on the event  $M^* \geq m_j$ ,  $M_j \leq m_j$ . Thus,  $M^* \geq m_j$  and  $M_j > m_j$  imply that either  $E$  or  $H$  does not occur and so  $\mathbb{P}(M_j > m_j \text{ and } M^* \geq m_j) \leq \mathbb{P}(\{E^c \cup H^c\} \cap \{j, j^* \in \mathcal{A}_{m_j}\}) \leq \mathbb{P}(E^c \cap j \in \mathcal{A}_{m_j}) + \mathbb{P}(H^c \cap j^* \in \mathcal{A}_{m_j})$ . Using Lemma 1, we then get that,

$$\mathbb{P}(M_j \geq m_j \text{ and } M^* \geq m_j) \leq \frac{6}{T\tilde{\Delta}_{m_j}^2}.$$

□

Note that the random set  $\mathcal{A}_m$  may not be defined for certain  $\omega \in \Omega$ . That is,  $\mathcal{A}_m$  is a partially defined random element. For convenience, we modify the definition of  $\mathcal{A}_m$  so that it is an emptyset for any  $\omega$  when it is not defined by the previous definition. Define the event  $F_j(m) = \{\bar{X}_{m,j^*} < \bar{X}_{m,j} - \tilde{\Delta}_m\} \cap \{j, j^* \in \mathcal{A}_m\}$  to be the event that arm  $j^*$  is eliminated by arm  $j$  in phase  $m$  (given our note on  $\mathcal{A}_m$ , this is well-defined). The probability of this occurring is bounded in the following lemma.

**Lemma 19** The probability that the optimal arm  $j^*$  is eliminated in round  $m < \infty$  by the suboptimal arm  $j$  is bounded by

$$\mathbb{P}(F_j(m)) \leq \frac{6}{T\tilde{\Delta}_m^2}.$$

*Proof:* First note that for a suboptimal arm  $j$  to eliminate arm  $j^*$  in round  $m$ , both  $j$  and  $j^*$  must be active in round  $m$  and  $\bar{X}_{m,j} - w_m > \bar{X}_{m,j^*} + w_m$ . Hence,

$$\mathbb{P}(F_j(m)) = \mathbb{P}(j, j^* \in \mathcal{A}_m \text{ and } \bar{X}_{m,j} - w_m > \bar{X}_{m,j^*} + w_m)$$

Then, observe that if

$$E = \{\bar{X}_{m,j} \leq \mu_j + w_m\} \quad \text{and} \quad H = \{\bar{X}_{m,j^*} > \mu^* - w_m\}$$

both hold in round  $m$ , it follows that,

$$\bar{X}_{m,j} - \tilde{\Delta}_m \leq \mu_j + w_m - \tilde{\Delta}_m \leq \mu_j - \frac{\tilde{\Delta}_m}{2} \leq \mu_{j^*} - \frac{\tilde{\Delta}_m}{2} \leq \bar{X}_{m,j^*} + w_m - \frac{\tilde{\Delta}_m}{2} \leq \bar{X}_{m,j^*}$$

so arm  $j^*$  will not be eliminated by arm  $j$  in round  $m$ . Hence, for arm  $j^*$  to be eliminated by arm  $j$  in round  $m$ , one of  $E$  or  $H$  must not occur and the probability of this is bounded by Lemma 1 as,

$$\mathbb{P}(F_j(m)) \leq \mathbb{P}((E^c \cup H^c) \cap (j, j^* \in \mathcal{A}_m)) \leq \mathbb{P}(E^c \cap (j \in \mathcal{A}_m)) + \mathbb{P}(H^c \cap (j^* \in \mathcal{A}_m)) \leq \frac{6}{T\tilde{\Delta}_m^2}.$$

□

We now return to bounding the expected regret in each of the two cases.

**Bounding Term I.** To bound the first term, we consider the cases where arm  $j$  is eliminated in or before the correct round ( $M_j \leq m_j$ ) and where arm  $j$  is eliminated late ( $M_j > m_j$ ). Then, by Lemma 18,

$$\begin{aligned}
 & \mathbb{E} \left[ \sum_{j=1}^K \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* \geq m_j\} \right] \\
 &= \mathbb{E} \left[ \sum_{j=1}^K \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* \geq m_j\} \mathbb{I}\{M_j \leq m_j\} \right] + \mathbb{E} \left[ \sum_{j=1}^K \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* \geq m_j\} \mathbb{I}\{M_j > m_j\} \right] \\
 &\leq \sum_{j=1}^K \mathbb{E}[\mathfrak{R}_T^{(j)} \mathbb{I}\{M_j \leq m_j\}] + \sum_{j=1}^K \mathbb{E}[T\Delta_j \mathbb{I}\{M^* \geq m_j, M_j > m_j\}] \\
 &\leq \sum_{j=1}^K \Delta_j n_{m_j} + \sum_{j=1}^K T\Delta_j \mathbb{P}(M_j > m_j \text{ and } M^* \geq m_j) \\
 &\leq \sum_{j=1}^K \Delta_j n_{m_j} + \sum_{j=1}^K T\Delta_j \frac{6}{T\tilde{\Delta}_2^{m_j}} \\
 &\leq \sum_{j=1}^K \left( \Delta_j n_{m_j} + \frac{24}{\tilde{\Delta}_2^{m_j}} \right) \leq \sum_{j=1}^K \left( \frac{96}{\Delta_j} + \Delta_j n_{m_j} \right).
 \end{aligned}$$

**Bounding Term II** For the second term, let  $m_{\max} = \max_{j \neq j^*} m_j$ . and recall that  $N_j$  is the total number of times arm  $j$  is played. Then,

$$\begin{aligned}
 \mathbb{E} \left[ \sum_{j=1}^K \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* < m_j\} \right] &= \mathbb{E} \left[ \sum_{m=1}^{m_{\max}} \sum_{j:m_j < m} \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\} \right] \\
 &= \sum_{m=1}^{m_{\max}} \mathbb{E} \left[ \mathbb{I}\{M^* = m\} \sum_{j:m_j > m} \mathfrak{R}_T^{(j)} \right] \\
 &= \sum_{m=1}^{m_{\max}} \mathbb{E} \left[ \mathbb{I}\{M^* = m\} \sum_{j:m_j > m} N_j \Delta_j \right] \\
 &\leq \sum_{m=1}^{m_{\max}} \mathbb{E} \left[ \mathbb{I}\{M^* = m\} T \max_{j:m_j > m} \Delta_j \right] \\
 &\leq \sum_{m=1}^{m_{\max}} 4\mathbb{P}(M^* = m) T \tilde{\Delta}_m.
 \end{aligned}$$

Now consider the probability that arm  $j^*$  is eliminated in round  $m$ . This includes the probability that it is eliminated by any suboptimal arm. For arm  $j^*$  to be eliminated in round  $m$  by a suboptimal arm with  $m_j < m$ , arm  $j$  must be active ( $M_j > m_j$ ) and the optimal arm must also have been active in round  $m_j$  ( $M^* \geq m_j$ ). Using this, it follows that

$$\begin{aligned}
 \mathbb{P}(M^* = m) &= \sum_{j=1}^K \mathbb{P}(F_j(m)) = \sum_{j:m_j < m} \mathbb{P}(F_j(m)) + \sum_{j:m_j \geq m} \mathbb{P}(F_j(m)) \\
 &\leq \sum_{j:m_j < m} \mathbb{P}(M_j > m_j \text{ and } M^* \geq m_j) + \sum_{j:m_j \geq m} \mathbb{P}(F_j(m)).
 \end{aligned}$$

Then, using Lemmas 18 and 19 and summing over all  $m \leq M$  gives,

$$\sum_{m=1}^{m_{\max}} \left( \sum_{j:m_j < m} 4\mathbb{P}(M_j > m_j \text{ and } M^* \geq m_j) T \tilde{\Delta}_m + \sum_{j:m_j \geq m} 4\mathbb{P}(F_j(m)) T \tilde{\Delta}_m \right)$$

$$\begin{aligned}
 &\leq \sum_{m=1}^{m_{\max}} \left( \sum_{j:m_j < m} 4 \frac{6}{T \tilde{\Delta}_m^2} T \frac{\tilde{\Delta}_{m_j}}{2^{m-m_j}} + \sum_{j:m_j \geq m} \frac{24}{T \tilde{\Delta}_m^2} T \tilde{\Delta}_m \right) \\
 &\leq \sum_{j=1}^K \frac{24}{\tilde{\Delta}_{m_j}} \sum_{m=m_j}^{m_{\max}} 2^{-(m-m_j)} + \sum_{j=1}^K \sum_{m=1}^{m_j} \frac{24}{2^{-m}} \\
 &\leq \sum_{j=1}^K \frac{96 \cdot 2}{\Delta_j} + \sum_{j=1}^K 24 \cdot 2^{m_j+1} \\
 &\leq \sum_{j=1}^K \frac{192}{\Delta_j} + \sum_{j=1}^K 48 \cdot \frac{4}{\Delta_j} = \sum_{j=1}^K \frac{384}{\Delta_j}.
 \end{aligned}$$

Combining the regret from terms I and II gives,

$$\mathbb{E}[\mathfrak{R}_T] \leq \sum_{j=1}^K \left( \frac{480}{\Delta_j} + \Delta_j n_{m_j} \right).$$

Hence, all that remains is to bound  $n_m$  in terms of  $\Delta_j, T$  and  $d$ ,

$$\begin{aligned}
 n_{m_j} &= \left\lceil \frac{1}{\tilde{\Delta}_{m_j}^2} \left( \sqrt{2 \log(T \tilde{\Delta}_{m_j}^2)} + \sqrt{2 \log(T \tilde{\Delta}_{m_j}^2) + \frac{8}{3} \tilde{\Delta}_{m_j} \log(T \tilde{\Delta}_{m_j}^2) + 6 \tilde{\Delta}_{m_j} m_j \mathbb{E}[\tau]} \right)^2 \right\rceil \\
 &\leq \left\lceil \frac{1}{\tilde{\Delta}_{m_j}^2} \left( 8 \log(T \tilde{\Delta}_{m_j}^2) + \frac{16}{3} \tilde{\Delta}_{m_j} \log(T \tilde{\Delta}_{m_j}^2) + 12 \tilde{\Delta}_{m_j} m_j \mathbb{E}[\tau] \right) \right\rceil \\
 &\leq 1 + \frac{8 \log(T \Delta_j^2/4)}{\tilde{\Delta}_{m_j}^2} + \frac{16 \log(T \Delta_j^2/4)}{3 \tilde{\Delta}_{m_j}} + \frac{12 \log_2(4/\Delta_j) \mathbb{E}[\tau]}{\tilde{\Delta}_{m_j}} \\
 &\leq 1 + \frac{128 \log(T \Delta_j^2)}{\Delta_j^2} + \frac{32 \log(T \Delta_j^2)}{3 \Delta_j} + \frac{96 \log(4/\Delta_j) \mathbb{E}[\tau]}{\Delta_j},
 \end{aligned}$$

where we have used  $(a+b)^2 \leq 2(a^2+b^2)$  for  $a, b \geq 0$  and  $\log_2(x) \leq 2 \log(x)$  for  $x > 0$ .

Hence, the total expected regret from ODAAF with bounded delays can be bounded by,

$$\mathbb{E}[\mathfrak{R}_t] \leq \sum_{j=1:j \neq j^*}^K \left( \frac{128 \log(T \Delta_j^2)}{\Delta_j} + \frac{32}{3} \log(T \Delta_j^2) + 96 \log(4/\Delta_j) \mathbb{E}[\tau] + \frac{480}{\Delta_j} + \Delta_j \right). \quad (16)$$

□

We now prove the problem independent regret bound,

**Corollary 3** *For any problem instance satisfying Assumption 1, the expected regret of Algorithm 1 satisfies*

$$\mathbb{E}[\mathfrak{R}_T] \leq O(\sqrt{KT \log(K)} + K \mathbb{E}[\tau] \log(T)).$$

*Proof:* Let

$$\lambda = \sqrt{\frac{K \log(K) e^2}{T}}$$

and note that for  $\Delta > \lambda$ ,  $\log(T \Delta^2)/\Delta$  is a decreasing function of  $\Delta$ . Then, for some constants  $C_1, C_2$ , and using the previous theorem, we can bound the regret by,

$$\mathbb{E}[\mathfrak{R}_T] \leq \sum_{j:\Delta_j \leq \lambda} \mathbb{E}[\mathfrak{R}_t^{(j)}] + \sum_{j:\Delta_j > \lambda} \mathbb{E}[\mathfrak{R}_T^{(j)}] \leq \frac{KC_1 \log(T \lambda^2)}{\lambda} + KdC_2 \log(1/\lambda) + T\lambda.$$

Then, substituting the above value of  $\lambda$  gives a worst case regret bound that scales with  $O(\sqrt{KT \log(K)} + K \mathbb{E}[\tau] \log(T))$ .

□

## C. Results for Delays with Bounded Support

### C.1. High Probability Bounds

**Lemma 5** *Under Assumptions 1 of known expected delay and 2 of bounded delays, and choice of  $n_m$  given in (6), the estimates  $\bar{X}_{m,j}$  obtained by Algorithm 1 satisfy the following: For any arm  $j$  and phase  $m$ , with probability at least  $1 - \frac{12}{T\bar{\Delta}_m^2}$ , either  $j \notin \mathcal{A}_m$  or*

$$\bar{X}_{m,j} - \mu_j \leq \bar{\Delta}_m/2.$$

*Proof:* Let

$$w_m = \frac{4 \log(T\bar{\Delta}_m^2)}{3n_m} + \sqrt{\frac{2 \log(T\bar{\Delta}_m^2)}{n_m}} + \frac{2\mathbb{E}[\tau]}{n_m}. \quad (17)$$

We show that with probability greater than  $1 - \frac{12}{T\bar{\Delta}_m^2}$ , either  $j \notin \mathcal{A}_m$  or  $\frac{1}{n_m} \sum_{t \in T_j(m)} (X_t - \mu_j) \leq w_m$ . For now, assume that  $n_m \geq md$ .

For arm  $j$  and phase  $m$ , assume  $j \in \mathcal{A}_m$  and define  $p_i$  to be the probability of the confidence bounds on arm  $j$  failing at the end of each phase  $i \leq m$ , ie.  $p_i \doteq \mathbb{P}(\sum_{t \in T_j(i)} (X_t - \mu_j) \geq n_i w_i)$  with  $p_0 = 0$ . Again, let  $B_{i,t} = R_t \mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\}$  and  $C_{i,t} = R_t \mathbb{I}\{\tau_{t,J_t} + t > U_{i,j}\}$  (note that we don't need to consider  $A_{i,t}$  since  $\nu_i = n_i - n_{i-1} \geq d$  so all reward entering  $[S_{i,j}, U_{i,j}]$  will be from the last  $\nu_i \geq d$  plays) and for any event  $H$ , let  $\mathbb{I}_i\{H\} := \mathbb{I}\{H \cap \{j \in \mathcal{A}_i\}\}$ . Recall the filtration  $\{\mathcal{G}_t\}_{t=0}^\infty$  from (12) where  $\mathcal{G}_t = \sigma(X_1, \dots, X_t, J_1, \dots, J_t, \tau_{1,J_1}, \dots, \tau_{t,J_t}, R_{1,J_1}, \dots, R_{t,J_t})$  and  $\mathcal{G}_0 = \{\emptyset, \Omega\}$ . Now, defining,

$$Q_t = \sum_{i=1}^m B_{i,t} \mathbb{I}\{S_{i,j} - d - 1 \leq t \leq S_{i,j} - 1\},$$

$$P_t = \sum_{i=1}^m C_{i,t} \mathbb{I}\{S_{i,j} \leq t \leq U_{i,j}\},$$

we use the decomposition

$$\begin{aligned} \sum_{t \in T_j(m)} (X_t - \mu_j) &= \sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} (X_t - \mu_j) \\ &\leq \sum_{i=1}^m \left( \sum_{t=S_{i-1,j}}^{S_{i,j}-1} R_{t,J_t} \mathbb{I}_i\{\tau_{t,J_t} + t \geq S_{i,j}\} + \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j) - \sum_{t=S_{i,j}}^{U_{i,j}} R_{t,J_t} \mathbb{I}_i\{\tau_{t,J_t} + t > U_{i,j}\} \right) \\ &\leq \sum_{i=1}^m \left( \sum_{t=S_{i,j}-d}^{S_{i,j}-1} B_{i,t} + \sum_{t=S_{i,j}}^{U_{i,j}} (R_t - \mu_j) - \sum_{t=S_{i,j}}^{U_{i,j}} C_{i,t} \right) \\ &= \sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j) + \sum_{t=1}^{S_{m,j}} Q_t - \sum_{t=1}^{U_{m,j}} P_t \\ &= \underbrace{\sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j)}_{\text{Term I.}} + \underbrace{\sum_{t=1}^{S_{m,j}} (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}])}_{\text{Term II.}} + \underbrace{\sum_{t=1}^{U_{m,j}} (\mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t)}_{\text{Term III.}} \\ &\quad + \underbrace{\sum_{t=1}^{S_{m,j}} \mathbb{E}[Q_t | \mathcal{G}_{t-1}] - \sum_{t=1}^{U_{m,j}} \mathbb{E}[P_t | \mathcal{G}_{t-1}]}_{\text{Term IV.}} \end{aligned}$$

**Outline of proof** Again, the proof continues by bounding each term of this decomposition in turn. Note that we do not have the  $A_{i,t}$  terms in this decomposition since there will be no reward from phase  $i - 1$  (before the bridge period)

received in  $[S_{i,j}, U_{i,j}]$ . We bound each of these terms with high probability. For terms I. and III., this is the same as in the general case (see the proof of Lemma 1, Appendix B). For term II. we need the following results to show that  $Z_t = Q_t - \mathbb{E}[Q_s | \mathcal{G}_{t-1}]$  is a martingale difference (Lemma 20) and to bound its variance (Lemma 21) before we can apply Freedman's inequality. The bound for term IV. is also different due to the bridge period and boundedness of the delay. After bounding each term, we collect them together and recursively calculate the probability with which the bounds hold.

**Lemma 20** *Let  $Y_s = \sum_{t=1}^s (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}])$  for all  $s \geq 1$ , and  $Y_0 = 0$ . Then  $\{Y_s\}_{s=0}^\infty$  is a martingale with respect to the filtration  $\{\mathcal{G}_s\}_{s=0}^\infty$  with increments  $Z_s = Y_s - Y_{s-1} = Q_s - \mathbb{E}[Q_s | \mathcal{G}_{s-1}]$  satisfying  $\mathbb{E}[Z_s | \mathcal{G}_{s-1}] = 0$ ,  $|Z_s| \leq 1$  for all  $s \geq 1$ .*

*Proof:* To show  $\{Y_s\}_{s=0}^\infty$  is a martingale we need to show that  $Y_s$  is  $\mathcal{G}_s$ -measurable for all  $s$  and  $\mathbb{E}[Y_s | \mathcal{G}_{s-1}] = Y_{s-1}$ .

Measurability: We show that  $B_{i,s} \mathbb{I}\{S_{i,j} - d - 1 \leq s \leq S_{i,j} - 1\}$  is  $\mathcal{G}_s$ -measurable. This then suffices to show that  $Y_s$  is  $\mathcal{G}_s$ -measurable since the filtration  $\mathcal{G}_s$  is non-decreasing in  $s$ .

First note that by definition of  $\mathcal{G}_s$ ,  $\tau_{t,J_t}$ ,  $R_{t,J_t}$  are all  $\mathcal{G}_s$ -measurable for  $t \leq s$ . Hence, it is sufficient to show that  $\mathbb{I}\{\tau_{s,J_s} + s \geq S_{i,j}, S_{i,j} - d - 1 \leq s \leq S_{i,j} - 1\}$  is  $\mathcal{G}_s$ -measurable since the product of measurable functions is measurable. For any  $s' \in \mathbb{N} \cup \{\infty\}$ ,  $\{S_{i,j} = s', s' - d - 1 \leq s\} \in \mathcal{G}_s$  for  $s \geq S_{i,j} - \nu_{i-1}$  and so the union  $\bigcup_{s' \in \mathbb{N} \cup \{\infty\}} \{\tau_{s,J_s} + s \geq s', s' - d - 1 \leq s \leq s' - 1, S_{i,j} = s'\} = \{\tau_{s,J_s} + s \geq S_{i,j}, S_{i,j} - d - 1 \leq s \leq S_{i,j} - 1\}$  is an element of  $\mathcal{G}_s$ .

Increments: Hence,  $\{Y_s\}_{s=0}^\infty$  is a martingale with respect to the filtration  $\{\mathcal{G}_s\}_{s=0}^\infty$  if the increments conditional on the past are zero. For any  $s \geq 1$ , we have that

$$Z_s = Y_s - Y_{s-1} = \sum_{t=1}^s (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]) - \sum_{t=1}^{s-1} (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]) = Q_s - \mathbb{E}[Q_s | \mathcal{G}_{s-1}].$$

Then,

$$\mathbb{E}[Z_s | \mathcal{G}_{s-1}] = \mathbb{E}[Q_s - \mathbb{E}[Q_s | \mathcal{G}_{s-1}] | \mathcal{G}_{s-1}] = \mathbb{E}[Q_s | \mathcal{G}_{s-1}] - \mathbb{E}[Q_s | \mathcal{G}_{s-1}] = 0$$

and so  $\{Y_s\}_{s=0}^\infty$  is a martingale.

Lastly, since for any  $t$  and  $\omega \in \Omega$ , there is at most one  $i$  where  $\mathbb{I}\{S_{i,j} - d \leq t \leq S_{i,j} - 1\}(\omega) = 1$ , and by definition of  $R_{t,J_t}$ ,  $B_{i,t} \leq 1$ , it follows that  $|Z_s| = |Q_s - \mathbb{E}[Q_s | \mathcal{G}_{s-1}]| \leq 1$  for all  $s$ .  $\square$

**Lemma 21** *For any  $t \geq 1$ , let  $Z_t = Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]$ , then*

$$\sum_{t=1}^{S_{m,j}-1} \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] \leq m \mathbb{E}[\tau].$$

*Proof:* Let us denote  $S' \doteq S_{m,j} - 1$ . Observe that

$$\sum_{t=1}^{S'} \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] = \sum_{t=1}^{S'} \mathbb{V}(Q_t | \mathcal{G}_{t-1}) \leq \sum_{t=1}^{S'} \mathbb{E}[Q_t^2 | \mathcal{G}_{t-1}] = \sum_{t=1}^{S'} \mathbb{E} \left[ \left( \sum_{i=1}^m (B_{i,t} \mathbb{I}\{S_{i,j} - d \leq t \leq S_{i,j} - 1\}) \right)^2 \middle| \mathcal{G}_{t-1} \right].$$

Then for all  $i = 1, \dots, m$ , all indicator terms  $\mathbb{I}\{S_{i,j} - d \leq t \leq S_{i,j} - 1\}$  are  $\mathcal{G}_{t-1}$ -measurable and only one can be non zero for any  $\omega \in \Omega$ . Hence, for any  $i, i' \leq m, i \neq i'$ ,

$$B_{i,t} \times \mathbb{I}\{S_{i,j} - d - 1 \leq t \leq S_{i,j} - 1\} \times B_{i',t} \times \mathbb{I}\{S_{i',j} - d - 1 \leq t \leq S_{i',j} - 1\} = 0,$$

Using the above we see that

$$\begin{aligned} \sum_{t=1}^{S'} \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] &\leq \sum_{t=1}^{S'} \mathbb{E} \left[ \left( B_{i,t} \mathbb{I}\{S_{i,j} - d - 1 \leq t \leq S_{i,j} - 1\} \right)^2 \middle| \mathcal{G}_{t-1} \right] \\ &= \sum_{t=1}^{S'} \mathbb{E} \left[ \sum_{i=1}^m B_{i,t}^2 \mathbb{I}\{S_{i,j} - d - 1 \leq t \leq S_{i,j} - 1\}^2 \middle| \mathcal{G}_{t-1} \right] \end{aligned}$$



$$\begin{aligned}
 &= \sum_{i=1}^m \sum_{t=1}^{S'_i} \mathbb{E}[B_{i,t}^2 \mathbb{I}\{S_{i,j} - d - 1 \leq t \leq S_{i,j} - 1\} | \mathcal{G}_{t-1}] \\
 &\quad \text{(using that the indicator is } \mathcal{G}_{t-1}\text{-measurable)} \\
 &\leq \sum_{i=1}^m \sum_{t=S_{i,j}-d-1}^{S_{i,j}-1} \mathbb{E}[B_{i,t}^2 | \mathcal{G}_{t-1}].
 \end{aligned}$$

Then, for any  $i \geq 1$ ,

$$\begin{aligned}
 \sum_{t=S_{i,j}-d-1}^{S_{i,j}-1} \mathbb{E}[B_{i,t}^2 | \mathcal{G}_{t-1}] &= \sum_{t=S_{i,j}-d-1}^{S_{i,j}-1} \mathbb{E}[R_{t,J_t}^2 \mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\} | \mathcal{G}_{t-1}] \\
 &\leq \sum_{t=S_{i,j}-d-1}^{S_{i,j}-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\} | \mathcal{G}_{t-1}] \\
 &= \sum_{s=0}^{\infty} \mathbb{I}\{S_{i,j} = s\} \sum_{t=s-d-1}^{s-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\} | \mathcal{G}_{t-1}] \\
 &= \sum_{s=0}^{\infty} \sum_{t=s-d-1}^{s-1} \mathbb{E}[\mathbb{I}\{S_{i,j} = s, \tau_{t,J_t} + t \geq S_{i,j}\} | \mathcal{G}_{t-1}] \\
 &\quad \text{(Since } S_{i,j} \geq S_i \text{ and so, due to the bridge period, } \{S_{i,j} = s\} \in \mathcal{G}_{t-1} \text{ for any } t \geq s - d) \\
 &= \sum_{s=0}^{\infty} \sum_{t=s-d-1}^{s-1} \mathbb{E}[\mathbb{I}\{S_{i,j} = s, \tau_{t,J_t} + t \geq s\} | \mathcal{G}_{t-1}] \\
 &= \sum_{s=0}^{\infty} \mathbb{I}\{S_{i,j} = s\} \sum_{t=s-d-1}^{s-1} \mathbb{P}(\tau_{t,J_t} + t \geq s) \\
 &\quad \text{(Since } \{S_{i,j} = s\} \in \mathcal{G}_{t-1} \text{ for any } t \geq s - d) \\
 &\leq \sum_{s=0}^{\infty} \mathbb{I}\{S_{i,j} = s\} \sum_{l=0}^{\infty} \mathbb{P}(\tau > l) \\
 &\leq \mathbb{E}[\tau].
 \end{aligned}$$

Combining all terms gives the result.  $\square$

We now return to bounding each term of the decomposition

**Bounding Term I:** For term II., as in Lemma 1, we can use Lemma 17 to get that with probability greater than  $1 - \frac{1}{T\tilde{\Delta}_m^2}$ ,

$$\sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j) \leq \sqrt{\frac{n_m \log(T\tilde{\Delta}_m^2)}{2}}.$$

**Bounding Term II:** For Term II., we will use Freedman's inequality (Theorem 10). From Lemma 20,  $\{Y_s\}_{s=0}^{\infty}$  with  $Y_s = \sum_{t=1}^s (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}])$  is a martingale with respect to  $\{\mathcal{G}_s\}_{s=0}^{\infty}$  with increments  $\{Z_s\}_{s=0}^{\infty}$  satisfying  $\mathbb{E}[Z_s | \mathcal{G}_{s-1}] = 0$  and  $Z_s \leq 1$  for all  $s$ . Further, by Lemma 21,  $\sum_{t=1}^s \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] \leq m\mathbb{E}[\tau] \leq \frac{4 \times 2^m \mathbb{E}[\tau]}{8} \leq n_m/8$  with probability 1. Hence we can apply Freedman's inequality to get that with probability greater than  $1 - \frac{1}{T\tilde{\Delta}_m^2}$ ,

$$\sum_{t=1}^{S_{m,j}} (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]) \leq \frac{2}{3} \log(T\tilde{\Delta}_m^2) + \sqrt{\frac{1}{8} n_m \log(T\tilde{\Delta}_m^2)}.$$

**Bounding Term III:** For Term III., we again use Freedman's inequality (Theorem 10). As in Lemma 1, we use Lemma 14 to show that  $\{Y'_s\}_{s=0}^\infty$  with  $Y'_s = \sum_{t=1}^s (\mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t)$  is a martingale with respect to  $\{\mathcal{G}_s\}_{s=0}^\infty$  with increments  $\{Z'_s\}_{s=0}^\infty$  satisfying  $\mathbb{E}[Z'_s | \mathcal{G}_{s-1}] = 0$  and  $Z'_s \leq 1$  for all  $s$ . Further, by Lemma 15,  $\sum_{t=1}^s \mathbb{E}[Z'_t | \mathcal{G}_{t-1}] \leq m\mathbb{E}[\tau] \leq n_m/8$  with probability 1. Hence, with probability greater than  $1 - \frac{1}{T\tilde{\Delta}_m^2}$ ,

$$\sum_{t=1}^{U_{m,j}} (\mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t) \leq \frac{2}{3} \log(T\tilde{\Delta}_m^2) + \sqrt{\frac{1}{8} n_m \log(T\tilde{\Delta}_m^2)}.$$

**Bounding Term IV:** For term IV., we consider the expected difference at each round  $1 \leq i \leq m$  and exploit the independence of  $\tau_{t,J_t}$  and  $R_{t,J_t}$ . Consider first  $i \geq 2$  and let  $j'_i$  be the arm played just before arm  $j$  is played in the  $i$ th phase (allowing for  $j'_i$  to be the last arm played in phase  $i-1$ ). Then, much in the same way as Lemma 21,

$$\begin{aligned} \sum_{t=S_{i,j}-d-1}^{S_{i,j}-1} \mathbb{E}[B_{i,t} | \mathcal{G}_{t-1}] &= \sum_{t=S_{i,j}-d-1}^{S_{i,j}-1} \mathbb{E}[R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\} | \mathcal{G}_{t-1}] \\ &= \sum_{s'=d+1}^\infty \sum_{s=s'}^\infty \mathbb{I}\{S_i = s', S_{i,j} = s\} \sum_{t=s-d}^{s-1} \mathbb{E}[R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\} | \mathcal{G}_{t-1}] \\ &= \sum_{s'=d+1}^\infty \sum_{s=s'}^\infty \sum_{t=s-d}^{s-1} \sum_{k=1}^K \mathbb{E}[R_{t,J_t} \mathbb{I}\{S_i = s', S_{i,j} = s, \tau_{t,J_t} + t \geq S_{i,j}, J_t = k\} | \mathcal{G}_{t-1}] \\ &\quad \text{(Due to the bridge period } \{S_i = s', S_{i,j} = s\} \in \mathcal{G}_{t-1} \text{ for } t \geq s-d \geq s'-d) \\ &= \sum_{s'=d+1}^\infty \sum_{s=s'}^\infty \sum_{t=s-d}^{s-1} \sum_{k=1}^K \mathbb{I}\{S_i = s', S_{i,j} = s, J_t = k\} \mathbb{E}[R_{t,k} \mathbb{I}\{\tau_{t,k} + t \geq s\} | \mathcal{G}_{t-1}] \\ &= \sum_{s'=d+1}^\infty \sum_{s=s'}^\infty \sum_{t=s-d}^{s-1} \sum_{k=1}^K \mu_k \mathbb{I}\{S_i = s', S_{i,j} = s, J_t = k\} \mathbb{P}(\tau \geq s-t) \\ &= \mu_{j'_i} \sum_{l=0}^{d-1} \mathbb{P}(\tau > l). \end{aligned}$$

A similar argument works for  $i=1, j>1$  with the simplification that  $S_{i,j}$  is not a random quantity but known. Finally, for  $i=1, j=1$  the sum is 0. Furthermore, using a similar argument, for all  $i, j$ ,

$$\begin{aligned} \sum_{t=S_{i,j}}^{U_{i,j}} \mathbb{E}[C_{i,t} | \mathcal{G}_{t-1}] &= \sum_{t=U_{i,j}-d+1}^{U_{i,j}} \mathbb{E}[C_{i,t} | \mathcal{G}_{t-1}] \\ &= \sum_{s'=d+1}^\infty \sum_{s=s'}^\infty \sum_{t=s-d}^s \mathbb{E}[R_{t,j} \mathbb{I}\{\tau_{t,j} + t > s\} \mathbb{I}\{U_{i,j} = s, S_i = s'\} | \mathcal{G}_{t-1}] \\ &= \mu_j \sum_{s=d+1}^\infty \mathbb{I}\{U_{i,j} = s, S_i = s'\} \sum_{t=s-d}^s \mathbb{P}(\tau + t > s) \\ &= \mu_j \sum_{l=0}^{d-1} \mathbb{P}(\tau > l). \end{aligned}$$

Combining these we get the following bound for term IV for all  $(i, j) \neq (1, 1)$ ,

$$\begin{aligned} \sum_{t=S_{i,j}-d-1}^{S_{i,j}-1} \mathbb{E}[B_{i,t} | \mathcal{G}_{t-1}] - \sum_{t=S_{i,j}}^{U_{i,j}} \mathbb{E}[C_{i,t} | \mathcal{G}_{t-1}] &\leq \mu_{j'_i} \sum_{l=0}^{d-1} \mathbb{P}(\tau > l) - \mu_j \sum_{l=0}^{d-1} \mathbb{P}(\tau > l) \\ &\leq |\mu_{j'_i} - \mu_j| \mathbb{E}[\tau]. \end{aligned}$$

If  $(i, j) = (1, 1)$  then we have the upper bounded by  $\mu_1 \mathbb{E}[\tau] \leq \mathbb{E}[\tau] = \tilde{\Delta}_0 \mathbb{E}[\tau]$  since no pay-off seeps in and we define  $\tilde{\Delta}_0 = 1$ .

Let  $p_i$  be the probability that the confidence bounds for one arm hold in phase  $i$  and  $p_0 = 0$ . Then, the probability that either arm  $j'_i$  or  $j$  is active in phase  $i$  when it should have been eliminated in or before phase  $i - 1$  is less than  $2p_{i-1}$ . If neither arm should have been eliminated by phase  $i$ , this means that their mean rewards are within  $\tilde{\Delta}_{i-1}$  of each other. Hence, with probability greater than  $1 - 2p_{i-1}$ ,

$$\sum_{t=S_{i,j}-d-1}^{S_{i,j}-1} \mathbb{E}[B_{i,t} | \mathcal{G}_{t-1}] - \sum_{t=S_{i,j}}^{U_{i,j}} \mathbb{E}[C_{i,t} | \mathcal{G}_{t-1}] \leq \tilde{\Delta}_{i-1} \mathbb{E}[\tau].$$

Then, summing over all phases gives that with probability greater than  $1 - 2 \sum_{i=0}^{m-1} p_i$ ,

$$\sum_{i=1}^m \left( \sum_{t=S_{i,j}-d-1}^{S_{i,j}-1} \mathbb{E}[B_{i,t} | \mathcal{G}_{t-1}] - \sum_{t=S_{i,j}}^{U_{i,j}} \mathbb{E}[C_{i,t} | \mathcal{G}_{t-1}] \right) \leq \mathbb{E}[\tau] \sum_{i=1}^m \tilde{\Delta}_{i-1} = \mathbb{E}[\tau] \sum_{i=0}^{m-1} \frac{1}{2^i} \leq 2\mathbb{E}[\tau].$$

**Combining all Terms:** To get the final high probability bound, we sum the bounds for each term I-IV.. Then, with probability greater than  $1 - (\frac{3}{T\tilde{\Delta}_m^2} + 2 \sum_{i=1}^{m-1} p_i)$  either  $j \notin \mathcal{A}_m$  or arm  $j$  is played  $n_m$  times by the end of phase  $m$  and

$$\begin{aligned} \frac{1}{n_m} \sum_{t \in T_j(m)} (X_t - \mu_j) &\leq \frac{4 \log(T\tilde{\Delta}_m^2)}{3n_m} + \left( \frac{2}{\sqrt{8}} + \frac{1}{\sqrt{2}} \right) \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{n_m}} + \frac{2\mathbb{E}[\tau]}{n_m} \\ &\leq \frac{4 \log(T\tilde{\Delta}_m^2)}{3n_m} + \sqrt{\frac{2 \log(T\tilde{\Delta}_m^2)}{n_m}} + \frac{2\mathbb{E}[\tau]}{n_m} = w_m. \end{aligned}$$

Using the fact that  $p_0 = 0$  and substituting the other  $p_i$ 's using the recursive relationship  $p_i = \frac{3}{T\tilde{\Delta}_i^2} + 2 \sum_{l=1}^{i-1} p_l$  gives,

$$\begin{aligned} \frac{3}{T\tilde{\Delta}_m^2} + 2 \sum_{i=0}^{m-1} p_i &= \frac{3}{T\tilde{\Delta}_m^2} + 2 \left( \frac{3}{T\tilde{\Delta}_{m-1}^2} + 2(p_{m-2} + \dots + p_1) + p_{m-2} + \dots + p_1 \right) \\ &= \frac{3}{T\tilde{\Delta}_m^2} + 2 \left( \frac{3}{T\tilde{\Delta}_{m-1}^2} + 3(p_{m-2} + \dots + p_1) \right) \\ &= \frac{3}{T\tilde{\Delta}_m^2} + 2 \left( \frac{3}{T\tilde{\Delta}_{m-1}^2} + 3 \left( \frac{3}{T\tilde{\Delta}_{m-2}^2} + 3(p_{m-3} + \dots + p_1) \right) \right) \\ &\leq \sum_{i=1}^m 3^{m-i} \frac{3}{T\tilde{\Delta}_i^2} \\ &= \frac{3}{T} \sum_{i=1}^m 3^{m-i} 2^{2i} \\ &= \frac{3}{T} \sum_{i=1}^m 3^{m-i} 4^i \\ &= \frac{3}{T} \sum_{i=1}^m \left( \frac{3}{4} \right)^{m-i} 4^{m-i} 4^i \\ &= \frac{3 \times 4^m}{T} \sum_{i=1}^m \left( \frac{3}{4} \right)^{m-i} \\ &\leq \frac{12}{T\tilde{\Delta}_m^2}. \end{aligned}$$

Hence, with probability greater than  $1 - \frac{12}{T\tilde{\Delta}_m^2}$ , either  $j \notin \mathcal{A}_m$  or  $\frac{1}{n_m} \sum_{t \in T_j(m)} (X_t - \mu_j) \leq w_m$ .

**Defining  $n_m$ :** The above results rely on the assumption that  $n_m \geq md$ , so that only the previous arm can corrupt our observations. In practice, if  $d$  is too large then we will not want to play each active arm  $d$  times per phase because we will end up playing sub-optimal arms too many times. In this case, it is better to ignore the bound on the delay and use the results from Lemma 1 to set  $n_m$  as in (14). Formalizing this gives

$$n_m = \max \left\{ m\tilde{d}_m, \left\lceil \frac{1}{\tilde{\Delta}_m^2} \left( \sqrt{2 \log(T\tilde{\Delta}_m^2)} + \sqrt{2 \log(T\tilde{\Delta}_m^2) + \frac{8}{3}\tilde{\Delta}_m \log(T\tilde{\Delta}_m^2) + 4\tilde{\Delta}_m \mathbb{E}[\tau]} \right)^2 \right\rceil \right\} \quad (18)$$

where  $\tilde{d}_m = \min\{d, \frac{(14)}{m}\}$ . This ensures that if  $d$  is small, we play each active arm enough times to ensure that  $w_m \leq \frac{\tilde{\Delta}_m}{2}$  for  $w_m$  in (17). Similarly, for large  $d$ , by Lemma 1, we know that  $n_m$  is sufficiently large to guarantee  $w_m \leq \frac{\tilde{\Delta}_m}{2}$  for  $w_m$  from (10).  $\square$

## C.2. Regret Bounds

We now prove the regret bound given in Theorem 6. Note that for these results, it is necessary to use the bridge period of the algorithm.

**Theorem 6** *Under Assumption 1 and bounded delay Assumption 2, the expected regret of Algorithm 1 satisfies*

$$\begin{aligned} \mathbb{E}[\mathfrak{R}_T] \leq & \sum_{j=1; j \neq j^*}^K O\left(\frac{\log(T\Delta_j^2)}{\Delta_j} + \mathbb{E}[\tau]\right) \\ & + \min\left\{d, \frac{\log(T\Delta_j^2)}{\Delta_j} + \log\left(\frac{1}{\Delta_j}\right)\mathbb{E}[\tau]\right\}. \end{aligned}$$

*Proof:* For any sub-optimal arm  $j$ , define  $M_j$  to be the random variable representing the phase arm  $j$  is eliminated in and note that if  $M_j$  is finite,  $j \in \mathcal{A}_{M_j}$  but  $j \notin \mathcal{A}_{M_j+1}$ . Then let  $m_j$  be the phase arm  $j$  should be eliminated in, that is  $m_j = \min\{m | \tilde{\Delta}_m < \frac{\Delta_j}{2}\}$  and note that, from the definition of  $\tilde{\Delta}_m$  in our algorithm, we get the relations

$$2^m = \frac{1}{\tilde{\Delta}_m}, \quad 2\tilde{\Delta}_{m_j} = \tilde{\Delta}_{m_j-1} \geq \frac{\Delta_j}{2} \quad \text{and so,} \quad \frac{\Delta_j}{4} \leq \tilde{\Delta}_{m_j} \leq \frac{\Delta_j}{2}. \quad (19)$$

Define  $\mathfrak{R}_T^{(j)}$  to be the regret contribution from each arm  $1 \leq j \leq K$  and let  $M^*$  be the round where the optimal arm  $j^*$  is eliminated. Hence,

$$\begin{aligned} \mathbb{E}[\mathfrak{R}_T] &= \mathbb{E}\left[\sum_{j=1}^K \mathfrak{R}_T^{(j)}\right] = \mathbb{E}\left[\sum_{m=0}^{\infty} \sum_{j=1}^K \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\}\right] \\ &= \mathbb{E}\left[\sum_{m=0}^{\infty} \sum_{j: m_j < m} \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\} + \sum_{j: m_j \geq m} \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\}\right] \\ &= \underbrace{\mathbb{E}\left[\sum_{m=0}^{\infty} \sum_{j: m_j < m} \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\}\right]}_{\text{I.}} + \underbrace{\mathbb{E}\left[\sum_{m=0}^{\infty} \sum_{j: m_j \geq m} \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\}\right]}_{\text{II.}} \end{aligned}$$

We will bound the regret in each of these cases in turn. First, however, we need the following results.

**Lemma 22** *For any suboptimal arm  $j$ , if  $j^* \in \mathcal{A}_{m_j}$ , then the probability arm  $j$  is not eliminated by round  $m_j$  is,*

$$\mathbb{P}(M_j > m_j \text{ and } M^* \geq m_j) \leq \frac{24}{T\tilde{\Delta}_{m_j}^2}$$

*Proof:* The proof is exactly that of Lemma 18 but using Lemma 5 to bound the probability of the confidence bounds on either arm  $j$  or  $j^*$  failing.  $\square$

Define the event  $F_j(m) = \{\bar{X}_{m, j^*} < \bar{X}_{m, j} - \tilde{\Delta}_m\} \cap \{j, j^* \in \mathcal{A}_m\}$  to be the event that arm  $j^*$  is eliminated by arm  $j$  in phase  $m$ . The probability of this occurring is bounded in the following lemma.

**Lemma 23** *The probability that the optimal arm  $j^*$  is eliminated in round  $m < \infty$  by the suboptimal arm  $j$  is bounded by*

$$\mathbb{P}(F_j(m)) \leq \frac{24}{T\tilde{\Delta}_m^2}$$

*Proof:* Again, the proof follows from Lemma 19 but using Lemma 5 to bound the probability of the confidence bounds failing.  $\square$

We now return to bounding the expected regret in each of the two cases.

**Bounding Term I.** To bound the first term, we consider the cases where arm  $j$  is eliminated in or before the correct round ( $M_j \leq m_j$ ) and where arm  $j$  is eliminated late ( $M_j > m_j$ ). Then,

$$\begin{aligned} \mathbb{E}\left[\sum_{m=0}^{\infty} \sum_{j:m_j < m} \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\}\right] &= \mathbb{E}\left[\sum_{j=1}^K \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* \geq m_j\}\right] \\ &= \mathbb{E}\left[\sum_{j=1}^K \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* \geq m_j\} \mathbb{I}\{M_j \leq m_j\}\right] + \mathbb{E}\left[\sum_{j=1}^K \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* \geq m_j\} \mathbb{I}\{M_j > m_j\}\right] \\ &\leq \sum_{j=1}^K \mathbb{E}[\mathfrak{R}_T^{(j)} \mathbb{I}\{M_j \leq m_j\}] + \sum_{j=1}^K \mathbb{E}[T\Delta_j \mathbb{I}\{M^* \geq m_j, M_j > m_j\}] \\ &\leq \sum_{j=1}^K 2\Delta_j n_{m_j, j} + \sum_{j=1}^K T\Delta_j \mathbb{P}(M_j > m_j \text{ and } M^* \geq m_j) \\ &\leq \sum_{j=1}^K 2\Delta_j n_{m_j, j} + \sum_{j=1}^K T\Delta_j \frac{24}{T\tilde{\Delta}_{m_j}^2} \\ &\leq \sum_{j=1}^K \left(2\Delta_j n_{m_j, j} + \frac{384}{\Delta_j}\right), \end{aligned}$$

where the extra factor of 2 comes from the fact that each arm will be played  $n_m$  times by the end of phase  $m$  to get the data for the estimated mean, then in the worst case, arm  $j$  is chosen as the arm to be played in the bridge period of each phase that it is active, and thus is played another  $n_m$  times.

**Bounding Term II** For the second term, we use the results from Theorem 2, but using Lemma 22 to bound the probability a suboptimal arm is eliminated in a later round and Lemma 23 to bound the probability  $j^*$  is eliminated by a suboptimal arm. Hence,

$$\mathbb{E}\left[\sum_{m=0}^{\infty} \sum_{j:m_j \geq m} \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\}\right] \leq \sum_{j=1}^K \frac{1536}{\Delta_j}.$$

Combining the regret from terms I and II gives,

$$\mathbb{E}[\mathfrak{R}_T] \leq \sum_{j=1}^K \left(\frac{1920}{\Delta_j} + 2\Delta_j n_{m_j, j}\right)$$

Hence, all that remains is to bound  $n_m$  in terms of  $\Delta_j, T$  and  $d$ . Using  $L_{m, T} = \log(T\tilde{\Delta}_m^2)$ , we have that,

$$\begin{aligned} n_{m_j, j} &= \max \left\{ m_j \tilde{d}_{m_j}, \left\lceil \frac{1}{\tilde{\Delta}_m^2} \left( \sqrt{2 \log(T\tilde{\Delta}_m)} + \sqrt{2 \log(T\tilde{\Delta}_m) + \frac{8}{3} \tilde{\Delta}_m \log(T\tilde{\Delta}_m) + 4\tilde{\Delta}_m \mathbb{E}[\tau]} \right)^2 \right\rceil \right\} \\ &\leq \max \left\{ m_j \tilde{d}_{m_j}, \left\lceil \frac{1}{\tilde{\Delta}_{m_j}^2} \left( 8L_{m_j, T} + \frac{16}{3} \tilde{\Delta}_{m_j} L_{m_j, T} + 8\tilde{\Delta}_{m_j} \mathbb{E}[\tau] \right) \right\rceil \right\} \end{aligned}$$

$$\begin{aligned} &\leq \max \left\{ m_j \tilde{d}_{m_j}, 1 + \frac{8L_{m_j, T}}{\tilde{\Delta}_{m_j}^2} + \frac{8L_{m_j, T}}{3\tilde{\Delta}_{m_j}} + \frac{8\mathbb{E}[\tau]}{\tilde{\Delta}_{m_j}} \right\} \\ &\leq \max \left\{ m_j \tilde{d}_{m_j}, 1 + \frac{128L_{m_j, T}}{\Delta_j^2} + \frac{32L_{m_j, T}}{\Delta_j} + \frac{32\mathbb{E}[\tau]}{\Delta_j} \right\} \end{aligned}$$

where we have used  $(a+b)^2 \leq 2(a^2+b^2)$  for  $a, b \geq 0$ .

Hence, using the definition of  $\tilde{d}_m = \min\{d, \frac{(14)}{m}\}$  and the results from Theorem 2, the total expected regret from ODAF with bounded delays can be bounded by,

$$\begin{aligned} \mathbb{E}[\mathfrak{R}_t] &\leq \sum_{j=1; j \neq j^*}^K \max \left\{ \min\{d, (16)\}, \left( \frac{256 \log(T\Delta_j^2)}{\Delta_j} + 64\mathbb{E}[\tau] + \frac{1920}{\Delta_j} + 64 \log(T\Delta_j^2) + 2\Delta_j \right) \right\}. \quad (20) \\ &\leq \sum_{j=1; j \neq j^*}^K \left( \frac{256 \log(T\Delta_j^2)}{\Delta_j} + 64\mathbb{E}[\tau] + \frac{1920}{\Delta_j} + 64 \log(T\Delta_j^2) + 2\Delta_j \right. \\ &\quad \left. + \min \left\{ d, \frac{128 \log(T\Delta_j^2)}{\Delta_j} + 96 \log(4/\Delta_j)\mathbb{E}[\tau] \right\} \right) \end{aligned}$$

□

Note that the constants in these regret bounds can be improved by only requiring the confidence bounds in phase  $m$  to hold with probability  $\frac{1}{T\Delta_m}$  rather than  $\frac{1}{T\Delta_m^2}$ . This comes at a cost of increasing the logarithmic term to  $\log(T\Delta_j)$ . We now prove the problem independent regret bound,

**Corollary 7** For any problem instance satisfying Assumptions 1 and 2 with  $d \leq \sqrt{\frac{T \log K}{K}} + \mathbb{E}[\tau]$ , the expected regret of Algorithm 1 satisfies

$$\mathbb{E}[\mathfrak{R}_T] \leq O(\sqrt{KT \log(K)} + K\mathbb{E}[\tau]).$$

*Proof:* We consider the maximal value each part of the regret in (20) can take. From Corollary 3, the first term is bounded by

$$O(\min\{Kd, \sqrt{KT \log K} + K \log(T)\mathbb{E}[\tau]\}).$$

For the first term, we again set  $\lambda = \sqrt{\frac{K \log(K)e^2}{T}}$ . Then, as in corollary Corollary 3, for constants  $C_1, C_2 > 0$ , we bound the regret contribution by

$$\sum_{j: \Delta_j \leq \lambda} \mathbb{E}[\mathfrak{R}_t^{(j)}] + \sum_{j: \Delta_j > \lambda} \mathbb{E}[\mathfrak{R}_t^{(j)}] \leq \frac{KC_1 \log(T\lambda^2)}{\lambda} + C_2 K\mathbb{E}[\tau] + T\lambda.$$

Then, substituting in for  $\lambda$  implies that the second term of (20) is  $O(\sqrt{KT \log K} + K\mathbb{E}[\tau])$ .

For  $d \leq \sqrt{\frac{T \log K}{K}} + \mathbb{E}[\tau]$ ,  $\min\{Kd, \sqrt{KT \log K} + K \log T\mathbb{E}[\tau]\} \leq \sqrt{KT \log K} + K\mathbb{E}[\tau]$ . Hence the bound in (20) gives

$$\mathbb{E}[\mathfrak{R}_T] \leq O(\sqrt{KT \log K} + K\mathbb{E}[\tau] + \sqrt{KT \log K} + K\mathbb{E}[\tau]) = O(\sqrt{KT \log K} + K\mathbb{E}[\tau]).$$

□

## D. Results for Delay with Known and Bounded Variance and Expectation

### D.1. High Probability Bounds

**Lemma 24** Under Assumption 1 of known expected value and 3 of known (bound on) the expectation and variance of the delay, and choice of  $n_m$  given in (7), the estimates  $\bar{X}_{m,j}$  obtained by Algorithm 1 satisfy the following: For any arm  $j$  and phase  $m$ , with probability at least  $1 - \frac{12}{T\Delta_m^2}$ , either  $j \notin \mathcal{A}_m$  or

$$\bar{X}_{m,j} - \mu_j \leq \tilde{\Delta}_m/2.$$

*Proof:* Let

$$w_m = \frac{4 \log(T \tilde{\Delta}_m^2)}{3n_m} + \sqrt{\frac{2 \log(T \tilde{\Delta}_m^2)}{n_m}} + \frac{2\mathbb{E}[\tau] + 4\mathbb{V}(\tau)}{n_m}. \quad (21)$$

We show that with probability greater than  $1 - \frac{12}{T \tilde{\Delta}_m^2}$ ,  $j \notin \mathcal{A}_m$  or  $\frac{1}{n_m} \sum_{t \in T_j(m)} (X_t - \mu_j) \leq w_m$ .

For any arm  $j$ , phase  $i$  and time  $t$ , define,

$$A_{i,t} = R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t \geq S_i\}, \quad B_{i,t} = R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\}, \quad C_{i,t} = R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t > U_{i,j}\} \quad (22)$$

as in (11) and

$$Q_t = \sum_{i=1}^m (A_{i,t} \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - \nu_{i-1} - 1\} + B_{i,t} \mathbb{I}\{S_i - \nu_{i-1} \leq t \leq S_{i,j} - 1\}),$$

$$P_t = \sum_{i=1}^m C_{i,t} \mathbb{I}\{S_{i,j} \leq t \leq U_{i,j}\},$$

where  $\nu_i = n_i - n_{i-1}$  is the number of times each active arm is played in phase  $i \geq 1$  (assume  $n_0 = 0$ ). Recall from the proof of Theorem 2,  $\mathbb{I}_i\{H\} := \mathbb{I}\{H \cap \{j \in \mathcal{A}_i\}\} \leq \mathbb{I}\{H\}$  and for all arms  $j$  and phases  $i$ ,  $\mathbb{I}_i\{\tau_{t,J_t} + t \geq S_{i,j}\} = \mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\}$  and  $\mathbb{I}_i\{\tau_{t,J_t} + t > U_{i,j}\} = \mathbb{I}\{\tau_{t,J_t} + t > U_{i,j}\}$ .

Then, using the convention  $S_0 = S_{0,j} = 0$  for all arms  $j$ , we use the decomposition,

$$\begin{aligned} \sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} (X_t - \mu_j) &\leq \sum_{i=1}^m \left( \sum_{t=S_{i-1,j}}^{S_{i,j}-1} R_{t,J_t} \mathbb{I}_i\{\tau_{t,J_t} + t \geq S_{i,j}\} + \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j) - \sum_{t=S_{i,j}}^{U_{i,j}} R_{t,J_t} \mathbb{I}_i\{\tau_{t,J_t} + t > U_{i,j}\} \right) \\ &\leq \sum_{i=1}^m \left( \sum_{t=S_{i-1,j}}^{S_i - \nu_{i-1} - 1} R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t \geq S_i\} + \sum_{t=S_i - \nu_{i-1}}^{S_{i,j}-1} R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\} \right. \\ &\quad \left. + \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j) - \sum_{t=S_{i,j}}^{U_{i,j}} R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t > U_{i,j}\} \right) \\ &= \sum_{i=1}^m \left( \sum_{t=S_{i-1,j}}^{S_i - \nu_{i-1} - 1} A_{i,t} + \sum_{t=S_i - \nu_{i-1}}^{S_{i,j}-1} B_{i,t} + \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j) - \sum_{t=S_{i,j}}^{U_{i,j}} C_{i,t} \right) \\ &= \sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j) + \sum_{t=1}^{S_{m,j}} Q_t - \sum_{t=1}^{U_{m,j}} P_t \\ &= \underbrace{\sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j)}_{\text{Term I.}} + \underbrace{\sum_{t=1}^{S_{m,j}} (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}])}_{\text{Term II.}} + \underbrace{\sum_{t=1}^{U_{m,j}} (\mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t)}_{\text{Term III.}} \\ &\quad + \underbrace{\sum_{t=1}^{S_{m,j}} \mathbb{E}[Q_t | \mathcal{G}_{t-1}] - \sum_{t=1}^{U_{m,j}} \mathbb{E}[P_t | \mathcal{G}_{t-1}]}_{\text{Term IV.}} \end{aligned} \quad (23)$$

Recall that the filtration  $\{\mathcal{G}_s\}_{s=0}^\infty$  is defined by  $\mathcal{G}_0 = \{\Omega, \emptyset\}$  and

$$\mathcal{G}_t = \sigma(X_1, \dots, X_t, J_1, \dots, J_t, \tau_{1,J_1}, \dots, \tau_{t,J_t}, R_{1,J_1}, \dots, R_{t,J_t}).$$

Furthermore, we have defined  $S_{i,j} = \infty$  if arm  $j$  is eliminated before phase  $i$  and  $S_i = \infty$  if the algorithm stops before reaching phase  $i$ .

**Outline of proof:** We will bound each term of the above decomposition in turn. We first show in Lemma 25 how the bounded second moment information can be incorporated using Chebychev's inequality. In Lemma 26, we show that  $Z_t = Q_t - \mathbb{E}[Q_t|\mathcal{G}_{t-1}]$  is a martingale difference sequence and bound its variance in Lemma 27 before using Freedman's inequality. Then in Lemma 28, we provide alternative (tighter) bounds on  $A_{i,t}, B_{i,t}, C_{i,t}$  which are used to bound term IV. All these results are then combined to give a high probability bound on the entire decomposition.

**Lemma 25** For any  $a > \lfloor \mathbb{E}[\tau] \rfloor + 1$ ,  $a \in \mathbb{N}$ ,

$$\sum_{l=a}^{\infty} \mathbb{P}(\tau \geq l) \leq \frac{\mathbb{V}(\tau)}{a - \lfloor \mathbb{E}[\tau] \rfloor - 1}.$$

*Proof:* For any  $b > a$ ,  $b \in \mathbb{N}$ , and by denoting  $\xi \doteq \lfloor \mathbb{E}(\tau) \rfloor$ ,

$$\begin{aligned} \sum_{l=a}^b \mathbb{P}(\tau \geq l) &= \sum_{l=a}^b \mathbb{P}(\tau - \xi \geq l - \xi) = \sum_{l=a-\xi}^{b-\xi} \mathbb{P}(\tau - \xi \geq l) \\ &\leq \sum_{l=a-\xi}^{b-\xi} \frac{\mathbb{V}(\tau)}{l^2} \quad (\text{by Chebychev's inequality since } l + \xi > \mathbb{E}[\tau] \text{ for } l \geq a - \xi) \\ &\leq \mathbb{V}(\tau) \sum_{l=a-\xi-1}^{b-\xi-1} \frac{1}{l(l+1)} \\ &= \mathbb{V}(\tau) \sum_{l=a-\xi-1}^{b-\xi-1} \left( \frac{1}{l} - \frac{1}{l+1} \right) \\ &= \mathbb{V}(\tau) \left( \frac{1}{a-\xi-1} - \frac{1}{b-\xi} \right). \end{aligned}$$

Hence, taking  $b \rightarrow \infty$  gives

$$\sum_{l=a}^{\infty} \mathbb{P}(\tau \geq l) \leq \mathbb{V}(\tau) \frac{1}{a - \xi - 1}.$$

□

**Lemma 26** Let  $Y_s = \sum_{t=1}^s (Q_t - \mathbb{E}[Q_t|\mathcal{G}_{t-1}])$  for all  $s \geq 1$ , and  $Y_0 = 0$ . Then  $\{Y_s\}_{s=0}^{\infty}$  is a martingale with respect to the filtration  $\{\mathcal{G}_s\}_{s=0}^{\infty}$  with increments  $Z_s = Y_s - Y_{s-1} = Q_s - \mathbb{E}[Q_s|\mathcal{G}_{s-1}]$  satisfying  $\mathbb{E}[Z_s|\mathcal{G}_{s-1}] = 0$ ,  $|Z_s| \leq 1$  for all  $s \geq 1$ .

*Proof:* To show  $\{Y_s\}_{s=0}^{\infty}$  is a martingale we need to show that  $Y_s$  is  $\mathcal{G}_s$ -measurable for all  $s$  and  $\mathbb{E}[Y_s|\mathcal{G}_{s-1}] = Y_{s-1}$ .

Measurability: We show that  $A_{i,s} \mathbb{I}\{S_{i-1,j} \leq s \leq S_i - \nu_{i-1}\} + B_{i,s} \mathbb{I}\{S_i - \nu_{i-1} + 1 \leq s \leq S_{i,j} - 1\}$  is  $\mathcal{G}_s$ -measurable for every  $i \leq m$ . This then suffices to show that  $Y_s$  is  $\mathcal{G}_s$ -measurable since each  $Q_t$  is a sum of such terms and the filtration  $\mathcal{G}_s$  is non-decreasing in  $s$ .

First note that by definition of  $\mathcal{G}_s$ ,  $\tau_{t,J_t}, R_{t,J_t}$  are all  $\mathcal{G}_s$ -measurable for  $t \leq s$ . It is sufficient to show that  $\mathbb{I}\{\tau_{s,J_s} + s \geq S_i, S_{i-1,j} \leq s \leq S_i - \nu_i\} + \mathbb{I}\{\tau_{s,J_s} + s \geq S_{i,j}, S_i - \nu_{i-1} + 1 \leq s \leq S_{i,j} - 1\}$  is  $\mathcal{G}_s$ -measurable since the product of measurable functions is measurable. The first summand is  $\mathcal{G}_s$  measurable since  $\{S_{i-1,j} \leq s\} \in \mathcal{G}_s$  and  $\{S_i = s', S_{i-1,j} \leq s\} \in \mathcal{G}_s$  for all  $s' \in \mathbb{N} \cup \{\infty\}$ . So the union  $\bigcup_{s' \in \mathbb{N} \cup \{\infty\}} \{\tau_{s,J_s} + s \geq s', S_{i-1,j} \leq s \leq s' - \nu_i, S_i = s'\} = \{\tau_{s,J_s} + s \geq S_i, S_{i-1,j} \leq s \leq S_i - \nu_{i-1}\}$  is an element of  $\mathcal{G}_s$ . The same argument works for the second summand since  $\{S_{i,j} = s', S_i - \nu_{i-1} + 1 \leq s\} \in \mathcal{G}_s$  for all  $s' \in \mathbb{N} \cup \{\infty\}$ .

Increments: Hence, to show that  $\{Y_s\}_{s=0}^{\infty}$  is a martingale with respect to the filtration  $\{\mathcal{G}_s\}_{s=0}^{\infty}$  it just remains to show that the increments conditional on the past are zero. For any  $s \geq 1$ , we have that

$$Z_s = Y_s - Y_{s-1} = \sum_{t=1}^s (Q_t - \mathbb{E}[Q_t|\mathcal{G}_{t-1}]) - \sum_{t=1}^{s-1} (Q_t - \mathbb{E}[Q_t|\mathcal{G}_{t-1}]) = Q_s - \mathbb{E}[Q_s|\mathcal{G}_{s-1}].$$



Then,

$$\mathbb{E}[Z_s | \mathcal{G}_{s-1}] = \mathbb{E}[Q_s - \mathbb{E}[Q_s | \mathcal{G}_{s-1}] | \mathcal{G}_{s-1}] = \mathbb{E}[Q_s | \mathcal{G}_{s-1}] - \mathbb{E}[Q_s | \mathcal{G}_{s-1}] = 0$$

and so  $\{Y_s\}_{s=0}^\infty$  is a martingale.

Lastly, since for any  $t$  and  $\omega \in \Omega$ , there is only one  $i$  where one of  $\mathbb{I}\{S_{i-1,j} \leq t \leq S_i - \nu_{i-1}\}$  or  $\mathbb{I}\{S_i - \nu_{i-1} + 1 \leq t \leq S_{i,j} - 1\}$  is equal to one (they cannot both be one), and by definition of  $R_{t,J_t}$ ,  $A_{i,t}, B_{i,t} \leq 1$ , it follows that  $|Z_s| = |Q_s - \mathbb{E}[Q_s | \mathcal{G}_{s-1}]| \leq 1$  for all  $s$ .  $\square$

**Lemma 27** For any  $t \geq 1$ , let  $Z_t = Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]$ , then

$$\sum_{t=1}^{S_{m,j}-1} \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] \leq m\mathbb{E}[\tau] + m\mathbb{V}(\tau).$$

*Proof:* Let us denote  $S' \doteq S_{m,j} - 1$ . Observe that

$$\begin{aligned} \sum_{t=1}^{S'} \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] &= \sum_{t=1}^{S'} \mathbb{V}(Q_t | \mathcal{G}_{t-1}) \leq \sum_{t=1}^{S'} \mathbb{E}[Q_t^2 | \mathcal{G}_{t-1}] \\ &= \sum_{t=1}^{S'} \mathbb{E} \left[ \left( \sum_{i=1}^m (A_{i,t} \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - \nu_{i-1} - 1\} + B_{i,t} \mathbb{I}\{S_i - \nu_{i-1} \leq t \leq S_{i,j} - 1\}) \right)^2 \middle| \mathcal{G}_{t-1} \right]. \end{aligned}$$

Then all indicator terms  $\mathbb{I}\{S_{i-1,j} \leq t \leq S_i - \nu_{i-1} - 1\}$  and  $\mathbb{I}\{S_i - \nu_{i-1} \leq t \leq S_{i,j} - 1\}$  for all  $i = 1, \dots, m$  are  $\mathcal{G}_{t-1}$ -measurable and only one can be non zero for any  $\omega \in \Omega$ . Hence, for any  $\omega \in \Omega$ , their product must be 0. Furthermore, for any  $i, i' \leq m, i \neq i'$ ,

$$\begin{aligned} A_{i,t} \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - \nu_{i-1} - 1\} \times A_{i',t} \mathbb{I}\{S_{i'-1,j} \leq t \leq S_{i'} - \nu_{i'-1} - 1\} &= 0, \\ B_{i,t} \mathbb{I}\{S_i - \nu_{i-1} \leq t \leq S_{i,j} - 1\} \times B_{i',t} \mathbb{I}\{S_{i'} - \nu_{i'-1} \leq t \leq S_{i',j} - 1\} &= 0, \\ A_{i,t} \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - \nu_{i-1} - 1\} \times B_{i',t} \mathbb{I}\{S_{i'} - \nu_{i'-1} \leq t \leq S_{i',j} - 1\} &= 0, \\ A_{i',t} \mathbb{I}\{S_{i'-1,j} \leq t \leq S_{i'} - \nu_{i'-1} - 1\} \times B_{i,t} \mathbb{I}\{S_i - \nu_{i-1} \leq t \leq S_{i,j} - 1\} &= 0. \end{aligned}$$

Using the above we see that,

$$\begin{aligned} \sum_{t=1}^{S'} \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] &\leq \sum_{t=1}^{S'} \mathbb{E} \left[ \left( \sum_{i=1}^m (A_{i,t} \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - \nu_{i-1} - 1\} + B_{i,t} \mathbb{I}\{S_i - \nu_{i-1} \leq t \leq S_{i,j} - 1\}) \right)^2 \middle| \mathcal{G}_{t-1} \right] \\ &= \sum_{t=1}^{S'} \mathbb{E} \left[ \sum_{i=1}^m (A_{i,t}^2 \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - \nu_{i-1} - 1\}^2 + B_{i,t}^2 \mathbb{I}\{S_i - \nu_{i-1} \leq t \leq S_{i,j} - 1\}^2) \middle| \mathcal{G}_{t-1} \right] \\ &= \sum_{i=2}^m \sum_{t=1}^{S'} \mathbb{E}[A_{i,t}^2 \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - \nu_{i-1} - 1\} | \mathcal{G}_{t-1}] \\ &\quad + \sum_{i=1}^m \sum_{t=1}^{S'} \mathbb{E}[B_{i,t}^2 \mathbb{I}\{S_i - \nu_{i-1} \leq t \leq S_{i,j} - 1\} | \mathcal{G}_{t-1}] \\ &\hspace{15em} \text{(using that both indicators are } \mathcal{G}_{t-1}\text{-measurable)} \\ &\leq \sum_{i=2}^m \sum_{t=S_{i-1,j}}^{S_i - \nu_{i-1} - 1} \mathbb{E}[A_{i,t}^2 | \mathcal{G}_{t-1}] + \sum_{i=1}^m \sum_{t=S_i - \nu_{i-1}}^{S_{i,j} - 1} \mathbb{E}[B_{i,t}^2 | \mathcal{G}_{t-1}]. \end{aligned}$$

Then, for any  $i \geq 2$ ,

$$\sum_{t=S_{i-1,j}}^{S_i - \nu_{i-1} - 1} \mathbb{E}[A_{i,t}^2 | \mathcal{G}_{t-1}] = \sum_{t=S_{i-1,j}}^{S_i - \nu_{i-1} - 1} \mathbb{E}[R_{t,J_t}^2 \mathbb{I}\{\tau_{t,J_t} + t \geq S_i\} | \mathcal{G}_{t-1}]$$

$$\begin{aligned}
 &\leq \sum_{t=S_{i-1,j}}^{S_i-\nu_{i-1}-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_i\} | \mathcal{G}_{t-1}] \\
 &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i-1,j} = s, S_i = s'\} \sum_{t=s}^{s'-\nu_{i-1}-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_i\} | \mathcal{G}_{t-1}] \\
 &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s}^{s'-\nu_{i-1}-1} \mathbb{E}[\mathbb{I}\{S_{i-1,j} = s, S_i = s', \tau_{t,J_t} + t \geq S_i\} | \mathcal{G}_{t-1}] \\
 &\hspace{15em} \text{(Since } \{S_i = s', S_{i-1,j} = s\} \in \mathcal{G}_{t-1} \text{ for } t \geq s) \\
 &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s}^{s'-\nu_{i-1}-1} \mathbb{E}[\mathbb{I}\{S_{i-1,j} = s, S_i = s', \tau_{t,J_t} + t \geq s'\} | \mathcal{G}_{t-1}] \\
 &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i-1,j} = s, S_i = s'\} \sum_{t=s}^{s'-\nu_{i-1}-1} \mathbb{P}(\tau_{t,J_t} + t \geq s') \\
 &\hspace{15em} \text{(Since } \{S_i = s', S_{i-1,j} = s\} \in \mathcal{G}_{t-1} \text{ for } t \geq s) \\
 &\leq \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i-1,j} = s, S_i = s'\} \sum_{l=\nu_{i-1}+1}^{\infty} \mathbb{P}(\tau > l) \\
 &\leq \mathbb{V}[\tau],
 \end{aligned}$$

by Lemma 25 since  $\nu_i \geq \lfloor \mathbb{E}[\tau] \rfloor + 2$  for all  $i$ . Likewise, for any  $i \geq 2$ ,

$$\begin{aligned}
 \sum_{t=S_i-\nu_{i-1}}^{S_{i,j}-1} \mathbb{E}[B_{i,t}^2 | \mathcal{G}_{t-1}] &= \sum_{t=S_i-\nu_{i-1}}^{S_{i,j}-1} \mathbb{E}[R_{t,J_t}^2 \mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\} | \mathcal{G}_{t-1}] \\
 &\leq \sum_{t=S_i-\nu_{i-1}}^{S_{i,j}-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\} | \mathcal{G}_{t-1}] \\
 &= \sum_{s=\nu_{i-1}+1}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_i = s, S_{i,j} = s'\} \sum_{t=s-\nu_{i-1}}^{s'-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\} | \mathcal{G}_{t-1}] \\
 &= \sum_{s=\nu_{i-1}+1}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s-\nu_{i-1}}^{s'-1} \mathbb{E}[\mathbb{I}\{S_i = s, S_{i,j} = s', \tau_{t,J_t} + t \geq s'\} | \mathcal{G}_{t-1}] \\
 &\hspace{15em} \text{(Since } \{S_{i,j} = s', S_i = s\} \in \mathcal{G}_{t-1} \text{ for } t \geq s - \nu_{i-1}) \\
 &= \sum_{s=\nu_{i-1}+1}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_i = s, S_{i,j} = s'\} \sum_{t=s-\nu_{i-1}}^{s'-1} \mathbb{P}(\tau_{t,J_t} + t \geq s') \\
 &\leq \sum_{s=\nu_{i-1}+1}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_i = s, S_{i,j} = s'\} \sum_{l=0}^{\infty} \mathbb{P}(\tau > l) \\
 &\leq \mathbb{E}[\tau]
 \end{aligned}$$

and for  $i = 1$  the derivation simplifies since we need to sum over 1 to  $S_{1,j} - 1$  only. Combining all terms gives the result.  $\square$

**Lemma 28** For  $A_{i,t}, B_{i,t}$  and  $C_{i,t}$  defined as in (22), let  $\nu_i = n_i - n_{i-1}$  be the number of times each arm is played in phase  $i$  and  $j'_i$  be the arm played directly before arm  $j$  in phase  $i$ . Then, it holds that, for any arm  $j$  and phase  $i \geq 1$ ,

$$(i) \quad \sum_{t=S_{i-1,j}}^{S_i-\nu_{i-1}-1} \mathbb{E}[A_{i,t} | \mathcal{G}_{t-1}] \leq \sum_{l=\nu_{i-1}+1}^{\infty} \mathbb{P}(\tau \geq l).$$

$$(ii) \sum_{t=S_i-\nu_{i-1}}^{S_{i,j}-1} \mathbb{E}[B_{i,t}|\mathcal{G}_{t-1}] \leq \sum_{l=\nu_{i-1}+1}^{\infty} \mathbb{P}(\tau \geq l) + \mu_{j'_i} \sum_{l=0}^{\nu_{i-1}} \mathbb{P}(\tau > l).$$

$$(iii) \sum_{t=S_{i,j}}^{U_{i,j}} \mathbb{E}[C_{i,t}|\mathcal{G}_{t-1}] = \mu_j \sum_{l=0}^{\nu_{i-1}} \mathbb{P}(\tau > l).$$

*Proof:* The proof is very similar to that of Lemma 27. We prove each statement individually.

**Statement (i):** This is similar to the proof of Lemma 27,

$$\begin{aligned} \sum_{t=S_{i-1,j}}^{S_i-\nu_{i-1}-1} \mathbb{E}[A_{i,t}|\mathcal{G}_{t-1}] &\leq \sum_{t=S_{i-1,j}}^{S_i-\nu_{i-1}-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_i\}|\mathcal{G}_{t-1}] \\ &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i-1,j} = s, S_i = s'\} \sum_{t=s}^{s'-\nu_{i-1}-1} \mathbb{E}[\mathbb{I}\{\tau_{t,J_t} + t \geq S_i\}|\mathcal{G}_{t-1}] \\ &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \sum_{t=s}^{s'-\nu_{i-1}-1} \mathbb{E}[\mathbb{I}\{S_{i-1,j} = s, S_i = s', \tau_{t,J_t} + t \geq s'\}|\mathcal{G}_{t-1}] \\ &\hspace{15em} (\text{Since } \{S_i = s', S_{i-1,j} = s\} \in \mathcal{G}_{t-1} \text{ for } t \geq s) \\ &= \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i-1,j} = s, S_i = s'\} \sum_{t=s}^{s'-\nu_{i-1}-1} \mathbb{P}(\tau_{t,J_t} + t \geq s') \\ &\leq \sum_{s=0}^{\infty} \sum_{s'=s}^{\infty} \mathbb{I}\{S_{i-1,j} = s, S_i = s'\} \sum_{l=\nu_{i-1}+1}^{\infty} \mathbb{P}(\tau > l) \\ &= \sum_{l=\nu_{i-1}+1}^{\infty} \mathbb{P}(\tau > l). \end{aligned}$$

**Statement (ii):** For statement (ii), we have that for  $(i, j) \neq (1, 1)$ ,

$$\sum_{t=S_i-\nu_{i-1}}^{S_{i,j}-1} \mathbb{E}[B_{i,t}|\mathcal{G}_{t-1}] = \sum_{t=S_i-\nu_{i-1}}^{S_{i,j}-\nu_{i-1}-2} \mathbb{E}[B_{i,t}|\mathcal{G}_{t-1}] + \sum_{t=S_{i,j}-\nu_{i-1}-1}^{S_{i,j}-1} \mathbb{E}[B_{i,t}|\mathcal{G}_{t-1}].$$

Then, since  $\{S_{i,j} = s'\} \cap \{S_i - \nu_{i-1} \leq t\} \in \mathcal{G}_{t-1}$  so we can use the same technique as for statement (i) to bound the first term. For the second term, since we will be playing only arm  $j'_i$  for  $S_{i,j} - \nu_{i-1} - 1, \dots, S_{i,j} - 1$ , so,

$$\begin{aligned} \sum_{t=S_{i,j}-\nu_{i-1}-1}^{S_{i,j}-1} \mathbb{E}[B_{i,t}|\mathcal{G}_{t-1}] &= \sum_{t=S_{i,j}-\nu_{i-1}-1}^{S_{i,j}-1} \mathbb{E}[R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\}|\mathcal{G}_{t-1}] \\ &= \sum_{s=0}^{\infty} \mathbb{I}\{S_{i,j} = s\} \sum_{t=s-\nu_{i-1}-1}^{s-1} \mathbb{E}[R_{t,J_t} \mathbb{I}\{\tau_{t,J_t} + t \geq S_{i,j}\}|\mathcal{G}_{t-1}] \\ &= \sum_{s=0}^{\infty} \sum_{t=s-\nu_{i-1}-1}^{s-1} \mathbb{E}[R_{t,J_t} \mathbb{I}\{S_{i,j} = s, \tau_{t,J_t} + t \geq S_{i,j}\}|\mathcal{G}_{t-1}] \\ &\hspace{15em} (\text{Since } \{S_{i,j} = s', S_{i,j} - \nu_{i-1} \leq t\} \in \mathcal{G}_{t-1}) \\ &= \sum_{s=0}^{\infty} \sum_{t=s-\nu_{i-1}-1}^{s-1} \mathbb{E}[R_{t,J_t} \mathbb{I}\{S_{i,j} = s, \tau_{t,J_t} + t \geq s\}|\mathcal{G}_{t-1}] \end{aligned}$$

$$\begin{aligned}
 &= \sum_{s=0}^{\infty} \mathbb{I}\{S_{i,j} = s\} \sum_{t=s-\nu_{i-1}-1}^{s-1} \mu_{j'_i} \mathbb{P}(\tau_{t,J_t} + t \geq s) \\
 &\text{(Since } \{S_{i,j} = s\} \in \mathcal{G}_{t-1} \text{ for } t \geq s - \nu_{i-1} - 1 \text{ and given } \mathcal{G}_{t-1}, R_{t,J_t} \text{ and } \tau_{t,J_t} \text{ are independent)} \\
 &= \sum_{s=0}^{\infty} \mathbb{I}\{S_{i,j} = s\} \mu_{j'_i} \sum_{l=0}^{\nu_{i-1}} \mathbb{P}(\tau > l) \\
 &= \mu_{j'_i} \sum_{l=0}^{\nu_{i-1}} \mathbb{P}(\tau > l).
 \end{aligned}$$

Then, for  $(i, j) = (1, 1)$ , the amount seeping in will be 0, so using  $\nu_0 = 0, \mu'_{1_1} = 0$ , the result trivially holds. Hence,

$$\sum_{t=S_i-\nu_{i-1}}^{S_{i,j}-1} \mathbb{E}[B_{i,t} | \mathcal{G}_{t-1}] \leq \sum_{l=\nu_{i-1}+1}^{\infty} \mathbb{P}(\tau \geq l) + \mu_{j'_i} \sum_{l=0}^{\nu_{i-1}} \mathbb{P}(\tau > l).$$

**Statement (iii):** This is the same as in Lemma 16. □

We now bound each term of the decomposition in (23).

**Bounding Term I.:** For Term I., we can again use Lemma 17 as in the proof of Lemma 1 to get that with probability greater than  $1 - \frac{1}{T\tilde{\Delta}_m^2}$ ,

$$\sum_{i=1}^m \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j) \leq \sqrt{\frac{n_m \log(T\tilde{\Delta}_m^2)}{2}}.$$

**Bounding Term II.:** For Term II., we will use Freedman's inequality (Theorem 10). From Lemma 26,  $\{Y_s\}_{s=0}^{\infty}$  with  $Y_s = \sum_{t=1}^s (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}])$  is a martingale with respect to  $\{\mathcal{G}_s\}_{s=0}^{\infty}$  with increments  $\{Z_s\}_{s=0}^{\infty}$  satisfying  $\mathbb{E}[Z_s | \mathcal{G}_{s-1}] = 0$  and  $Z_s \leq 1$  for all  $s$ . Further, by Lemma 27,  $\sum_{t=1}^s \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] \leq m\mathbb{E}[\tau] + m\mathbb{V}(\tau) \leq \frac{4 \times 2^m}{8} (\mathbb{E}[\tau] + \mathbb{V}(\tau)) \leq n_m/8$  with probability 1. Hence we can apply Freedman's inequality to get that with probability greater than  $1 - \frac{1}{T\tilde{\Delta}_m^2}$ ,

$$\sum_{t=1}^{S_{m,j}} (Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]) = \sum_{s=1}^{\infty} \mathbb{I}\{S_{m,j} = s\} \times Y_s \leq \frac{2}{3} \log(T\tilde{\Delta}_m^2) + \sqrt{\frac{1}{8} n_m \log(T\tilde{\Delta}_m^2)},$$

using that Freedman's inequality applies simultaneously to all  $s \geq 1$ .

**Bounding Term III.:** For Term III., we again use Freedman's inequality (Theorem 10), using Lemma 14 to show that  $\{Y'_s\}_{s=0}^{\infty}$  with  $Y'_s = \sum_{t=1}^s (\mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t)$  is a martingale with respect to  $\{\mathcal{G}_s\}_{s=0}^{\infty}$  with increments  $\{Z'_s\}_{s=0}^{\infty}$  satisfying  $\mathbb{E}[Z'_s | \mathcal{G}_{s-1}] = 0$  and  $Z'_s \leq 1$  for all  $s$ . Further, by Lemma 15,  $\sum_{t=1}^s \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] \leq m\mathbb{E}[\tau] \leq n_m/8$  with probability 1. Hence, with probability greater than  $1 - \frac{1}{T\tilde{\Delta}_m^2}$ ,

$$\sum_{t=1}^{U_{m,j}} (\mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t) = \sum_{s=1}^{\infty} \mathbb{I}\{U_{m,j} = s\} \times Y'_s \leq \frac{2}{3} \log(T\tilde{\Delta}_m^2) + \sqrt{\frac{1}{8} n_m \log(T\tilde{\Delta}_m^2)}.$$

**Bounding Term IV.:** To begin with, observe that,

$$\begin{aligned}
 &\sum_{t=1}^{S_{m,j}} \mathbb{E}[Q_t | \mathcal{G}_{t-1}] - \sum_{t=1}^{U_{m,j}} \mathbb{E}[P_t | \mathcal{G}_{t-1}] \\
 &= \sum_{t=1}^{S_{m,j}} \mathbb{E} \left[ \sum_{i=1}^m (A_{i,t} \times \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - \nu_{i-1} - 1\} + B_{i,t} \times \mathbb{I}\{S_i - \nu_{i-1} \leq t \leq S_{i,j} - 1\}) \middle| \mathcal{G}_{t-1} \right]
 \end{aligned}$$

$$\begin{aligned}
 & - \sum_{t=1}^{U_{m,j}} \mathbb{E} \left[ \sum_{i=1}^m C_{i,t} \times \mathbb{I}\{S_{i,j} \leq t \leq U_{i,j}\} \middle| \mathcal{G}_{t-1} \right] \\
 = & \sum_{i=1}^m \sum_{t=1}^{S_{m,j}} \mathbb{E}[A_{i,t} \times \mathbb{I}\{S_{i-1,j} \leq t \leq S_i - \nu_{i-1} - 1\} | \mathcal{G}_{t-1}] \\
 & + \sum_{i=1}^m \sum_{t=1}^{S_{m,j}} \mathbb{E}[B_{i,t} \times \mathbb{I}\{S_i - \nu_{i-1} \leq t \leq S_{i,j} - 1\} | \mathcal{G}_{t-1}] \\
 & - \sum_{i=1}^m \sum_{t=1}^{U_{m,j}} \mathbb{E}[C_{i,t} \times \mathbb{I}\{S_{i,j} \leq t \leq U_{i,j}\} | \mathcal{G}_{t-1}] \\
 = & \sum_{i=1}^m \left( \sum_{t=S_{i-1,j}}^{S_i - \nu_{i-1} - 1} \mathbb{E}[A_{i,t} | \mathcal{G}_{t-1}] + \sum_{t=S_i - \nu_{i-1}}^{S_{i,j} - 1} \mathbb{E}[B_{i,t} | \mathcal{G}_{t-1}] - \sum_{t=S_{i,j}}^{U_{i,j}} \mathbb{E}[C_{i,t} | \mathcal{G}_{t-1}] \right) \\
 & \hspace{15em} \text{(using that the indicators are } \mathcal{G}_{t-1}\text{-measurable)} \\
 \leq & \sum_{i=1}^m \left( \sum_{l=\nu_{i-1}+1}^{\infty} \mathbb{P}(\tau \geq l) + \mu_{j'_i} \sum_{l=0}^{\nu_{i-1}} \mathbb{P}(\tau > l) - \mu_j \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l) \right), \\
 \leq & \sum_{i=1}^m \left( \frac{2\mathbb{V}(\tau)}{\nu_{i-1} - \mathbb{E}[\tau]} + (\mu_{j'_i} - \mu_j) \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l) \right), \\
 \leq & \sum_{i=1}^m \left( \frac{2\mathbb{V}(\tau)}{2^{i-1}} + (\mu_{j'_i} - \mu_j) \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l) \right), \tag{24}
 \end{aligned}$$

by Lemma 28 and Lemma 25 where we have used the fact that since  $n_m \leq T$ , the maximal number of rounds of the algorithm is  $\frac{1}{2} \log_2(T/4)$  and for  $m \leq \frac{1}{2} \log_2(T/4)$ ,  $\frac{\log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} \geq \frac{2 \log(T\tilde{\Delta}_{m-1}^2)}{\tilde{\Delta}_{m-1}^2}$  so  $n_m \geq 2n_{m-1}$  and  $\nu_m \geq n_{m-1}$ . Then for  $\mathbb{E}[\tau] \geq 1$ ,  $\nu_{i-1} - \mathbb{E}[\tau] \geq 2/\tilde{\Delta}_{i-1}\mathbb{E}[\tau] - \mathbb{E}[\tau] \geq (2 \times 2^{i-1} - 1)\mathbb{E}[\tau] \geq 2^{i-1}\mathbb{E}[\tau] \geq 2^{i-1}$  and for  $\mathbb{E}[\tau] \leq 1$ ,  $\nu_{i-1} - \mathbb{E}[\tau] \geq \nu_{i-1} - 1 \geq 2 \log(4)/\tilde{\Delta}_{i-1} - 1 \geq 2^{i-1}$  so  $\nu_{i-1} - \mathbb{E}[\tau] \geq 2^{i-1}$ . Then, the probability that either arm  $j'_i$  or  $j$  is active in phase  $i$  when it should have been eliminated in or before phase  $i-1$  is less than  $2p_{i-1}$ , where  $p_i$  is the probability that the confidence bounds for one arm holds in phase  $i$  and  $p_0 = 0$ . If neither arm should have been eliminated by phase  $i$ , this means that their mean rewards are within  $\tilde{\Delta}_{i-1}$  of each other. Hence, with probability greater than  $1 - 2p_{i-1}$ ,

$$\mu_{j'_i} \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l) - \mu_j \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l) \leq \tilde{\Delta}_{i-1} \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l) \leq \tilde{\Delta}_{i-1} \mathbb{E}[\tau].$$

Then, summing over all phases gives that with probability greater than  $1 - 2 \sum_{i=0}^{m-1} p_i$ ,

$$\begin{aligned}
 \sum_{t=1}^{S_{m,j}} \mathbb{E}[Q_t | \mathcal{G}_{t-1}] - \sum_{t=1}^{U_{m,j}} \mathbb{E}[P_t | \mathcal{G}_{t-1}] & \leq 2\mathbb{V}(\tau) \sum_{i=1}^m \frac{1}{2^{i-1}} + \mathbb{E}[\tau] \sum_{i=1}^m \tilde{\Delta}_{i-1} = (2\mathbb{V}(\tau) + \mathbb{E}[\tau]) \sum_{i=0}^{m-1} \frac{1}{2^i} \\
 & \leq 4\mathbb{V}(\tau) + 2\mathbb{E}[\tau].
 \end{aligned}$$

**Combining all terms:** To get the final high probability bound, we sum the bounds for each term I-IV.. Then, with probability greater than  $1 - (\frac{3}{T\tilde{\Delta}_m^2} + 2 \sum_{i=1}^{m-1} p_i)$ , either  $j \notin \mathcal{A}_m$  or arm  $j$  is played  $n_m$  times by the end of phase  $m$  and

$$\begin{aligned}
 \frac{1}{n_m} \sum_{t \in T_j(m)} (X_t - \mu_j) & \leq \frac{4 \log(T\tilde{\Delta}_m^2)}{3n_m} + \left( \frac{2}{\sqrt{8}} + \frac{1}{\sqrt{2}} \right) \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{n_m}} + \frac{2\mathbb{E}[\tau] + 4\mathbb{V}(\tau)}{n_m} \\
 & \leq \frac{4 \log(T\tilde{\Delta}_m^2)}{3n_m} + \sqrt{\frac{2 \log(T\tilde{\Delta}_m^2)}{n_m}} + \frac{2\mathbb{E}[\tau] + 4\mathbb{V}(\tau)}{n_m} = w_m.
 \end{aligned}$$

Using the fact that  $p_0 = 0$  and substituting the other  $p_i$ 's using the same recursive relationship  $p_i = \frac{3}{T\tilde{\Delta}_i^2} + 2\sum_{l=1}^{i-1} p_l$  as in the case for bounded delays (see the proof of Lemma 5) gives,  $p_m = \frac{12}{T\tilde{\Delta}_m^2}$  so the above bound holds with probability greater than  $1 - \frac{12}{T\tilde{\Delta}_m^2}$ .

**Defining  $n_m$ :** Setting

$$n_m = \left\lceil \frac{1}{\tilde{\Delta}_m^2} \left( \sqrt{2\log(T\tilde{\Delta}_m^2)} + \sqrt{2\log(T\tilde{\Delta}_m^2) + \frac{8}{3}\tilde{\Delta}_m \log(T\tilde{\Delta}_m^2) + 4\tilde{\Delta}_m(\mathbb{E}[\tau] + 2\mathbb{V}(\tau))} \right)^2 \right\rceil. \quad (25)$$

ensures that  $w_m \leq \frac{\tilde{\Delta}_m}{2}$  which concludes the proof.  $\square$

**Remark:** Note that if  $\mathbb{E}[\tau] \geq 1$ , then the confidence bounds can be tightened by replacing (24) with

$$\sum_{i=1}^m \left( \frac{2\mathbb{V}(\tau)}{2^{i-1}\mathbb{E}[\tau]} + (\mu_{j'_i} - \mu_j) \sum_{l=0}^{\nu_i} \mathbb{P}(\tau > l) \right)$$

This is obtained by noting that for  $\mathbb{E}[\tau] \geq 1$ ,  $\nu_{i-1} - \mathbb{E}[\tau] \geq 2/\tilde{\Delta}_{i-1}\mathbb{E}[\tau] - \mathbb{E}[\tau] \geq (2 \times 2^{i-1} - 1)\mathbb{E}[\tau] \geq 2^{i-1}\mathbb{E}[\tau]$ . This leads to replacing the  $\mathbb{V}(\tau)$  term in the definition of  $n_m$  by  $\mathbb{V}(\tau)/\mathbb{E}[\tau]$ .

## D.2. Regret Bounds

**Theorem 8** Under Assumption 1 and Assumption 3 of known (bound on) the expectation and variance of the delay, and choice of  $n_m$  from (7), the expected regret of Algorithm 1 can be upper bounded by,

$$\mathbb{E}[\mathfrak{R}_T] \leq \sum_{j=1: \mu_j \neq \mu^*}^K O\left(\frac{\log(T\Delta_j^2)}{\Delta_j} + \mathbb{E}[\tau] + \mathbb{V}(\tau)\right).$$

*Proof:* The proof is very similar to that of Theorem 2, however, for clarity, we repeat the main arguments here. For any sub-optimal arm  $j$ , define  $M_j$  to be the random variable representing the phase arm  $j$  is eliminated in and note that if  $M_j$  is finite,  $j \in \mathcal{A}_{M_j}$  but  $j \notin \mathcal{A}_{M_j+1}$ . Then let  $m_j$  be the phase arm  $j$  should be eliminated in, that is  $m_j = \min\{m | \tilde{\Delta}_m < \frac{\Delta_j}{2}\}$  and note that, from the new definition of  $\tilde{\Delta}_m$  in our algorithm, we get the relations

$$2^m = \frac{1}{\tilde{\Delta}_m}, \quad 2\tilde{\Delta}_{m_j} = \tilde{\Delta}_{m_j-1} \geq \frac{\Delta_j}{2} \quad \text{and so,} \quad \frac{\Delta_j}{4} \leq \tilde{\Delta}_{m_j} \leq \frac{\Delta_j}{2}. \quad (26)$$

Define  $\mathfrak{R}_T^{(j)}$  to be the regret contribution from each arm  $1 \leq j \leq K$  and let  $M^*$  be the round where the optimal arm  $j^*$  is eliminated. Hence,

$$\begin{aligned} \mathbb{E}[\mathfrak{R}_T] &= \mathbb{E}\left[\sum_{j=1}^K \mathfrak{R}_T^{(j)}\right] = \mathbb{E}\left[\sum_{m=0}^{\infty} \sum_{j=1}^K \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\}\right] \\ &= \mathbb{E}\left[\sum_{m=0}^{\infty} \sum_{j: m_j < m} \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\} + \sum_{j: m_j \geq m} \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\}\right] \\ &= \underbrace{\mathbb{E}\left[\sum_{m=0}^{\infty} \sum_{j: m_j < m} \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\}\right]}_{\text{i.}} + \underbrace{\mathbb{E}\left[\sum_{m=0}^{\infty} \sum_{j: m_j \geq m} \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\}\right]}_{\text{ii.}} \end{aligned}$$

We will bound the regret in each of these cases in turn. First, however, we need the following results.

**Lemma 29** For any suboptimal arm  $j$ , if  $j^* \in \mathcal{A}_{m_j}$ , then the probability arm  $j$  is not eliminated by round  $m_j$  is,

$$\mathbb{P}(M_j > m_j \text{ and } M^* \geq m_j) \leq \frac{24}{T\tilde{\Delta}_{m_j}^2}$$

*Proof:* The proof is exactly that of Lemma 18 but using Lemma 24 to bound the probability of the confidence bounds on either arm  $j$  or  $j^*$  failing.  $\square$

Define the event  $F_j(m) = \{\bar{X}_{m,j^*} < \bar{X}_{m,j} - \tilde{\Delta}_m\} \cap \{j, j^* \in \mathcal{A}_m\}$  to be the event that arm  $j^*$  is eliminated by arm  $j$  in phase  $m$ . The probability of this event is bounded in the following lemma.

**Lemma 30** *The probability that the optimal arm  $j^*$  is eliminated in round  $m < \infty$  by the suboptimal arm  $j$  is bounded by*

$$\mathbb{P}(F_j(m)) \leq \frac{24}{T\tilde{\Delta}_m^2}$$

*Proof:* Again, the proof follows from Lemma 19 but using Lemma 24 to bound the probability of the confidence bounds failing.  $\square$

We now return to bounding the expected regret in each of the two cases.

**Bounding Term I.** As in the proof of Theorem 2, to bound the first term, we consider the cases where arm  $j$  is eliminated in or before the correct round ( $M_j \leq m_j$ ) and where arm  $j$  is eliminated late ( $M_j > m_j$ ). Then, using Lemma 22,

$$\mathbb{E} \left[ \sum_{m=0}^{\infty} \sum_{j:m_j < m} \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\} \right] \leq \sum_{j=1}^K \left( 2\Delta_j n_{m_j, j} + \frac{384}{\Delta_j} \right)$$

**Bounding Term II** For the second term, we again use the results from Theorem 2, but using Lemma 29 to bound the probability a suboptimal arm is eliminated in a later round and Lemma 30 to bound the probability  $j^*$  is eliminated by a suboptimal arm. Hence,

$$\mathbb{E} \left[ \sum_{m=0}^{\infty} \sum_{j:m_j \geq m} \mathfrak{R}_T^{(j)} \mathbb{I}\{M^* = m\} \right] \leq \sum_{j=1}^K \frac{1920}{\Delta_j}.$$

Combining the regret from terms I and II gives,

$$\mathbb{E}[\mathfrak{R}_T] \leq \sum_{j=1}^K \left( \frac{1920}{\Delta_j} + 2\Delta_j n_{m_j, j} \right)$$

Hence, all that remains is to bound  $n_m$  in terms of  $\Delta_j, T$  and  $\mathbb{E}[\tau], \mathbb{V}(\tau)$ . Using  $L_{m,T} = \log(T\tilde{\Delta}_m^2)$ , we have that,

$$\begin{aligned} n_{m_j, j} &= \left\lceil \frac{1}{\tilde{\Delta}_m^2} \left( \sqrt{2 \log(T\tilde{\Delta}_m^2)} + \sqrt{2 \log(T\tilde{\Delta}_m^2) + \frac{8}{3} \tilde{\Delta}_m \log(T\tilde{\Delta}_m) + 4\tilde{\Delta}_m (\mathbb{E}[\tau] + 2\mathbb{V}(\tau))} \right)^2 \right\rceil \\ &\leq \left\lceil \frac{1}{\tilde{\Delta}_{m_j}^2} \left( 8L_{m_j, T} + \frac{16}{3} \tilde{\Delta}_{m_j} L_{m_j, T} + 8\tilde{\Delta}_{m_j} \mathbb{E}[\tau] + 16\tilde{\Delta}_{m_j} \mathbb{V}(\tau) \right) \right\rceil \\ &\leq 1 + \frac{8L_{m_j, T}}{\tilde{\Delta}_{m_j}^2} + \frac{16L_{m_j, T}}{3\tilde{\Delta}_{m_j}} + \frac{8\mathbb{E}[\tau]}{\tilde{\Delta}_{m_j}} + \frac{16\mathbb{V}(\tau)}{\tilde{\Delta}_{m_j}} \\ &\leq 1 + \frac{128L_{m_j, T}}{\Delta_j^2} + \frac{32L_{m_j, T}}{\Delta_j} + \frac{32\mathbb{E}[\tau]}{\Delta_j} + \frac{64\mathbb{V}(\tau)}{\Delta_j}. \end{aligned}$$

where we have used  $(a+b)^2 \leq 2(a^2 + b^2)$  for  $a, b \geq 0$ .

Hence, the total expected regret from ODAAF with bounded delays can be bounded by,

$$\mathbb{E}[\mathfrak{R}_t] \leq \sum_{j=1}^K \left( \frac{256 \log(T\Delta_j^2)}{\Delta_j} + 64\mathbb{E}[\tau] + 128\mathbb{V}(\tau) + \frac{1920}{\Delta_j} + 64 \log(T) + 2\Delta_j \right).$$

$\square$

Note that again, these constants can be improved at a cost of increasing  $\log(T\Delta_j^2)$  to  $\log(T\Delta_j)$ . We now prove the problem independent regret bound.

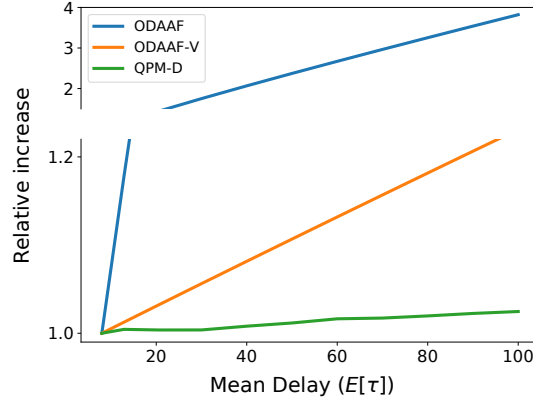


Figure 5: The relative increase in regret at horizon  $T = 250000$  for increasing mean delay when the delay is  $\mathcal{N}_+$  with variance 100.

**Corollary 9** For any problem instance satisfying Assumptions 1 and 3, the expected regret of Algorithm 1 satisfies

$$\mathbb{E}[\mathfrak{R}_T] \leq O(\sqrt{KT \log(K)} + K\mathbb{E}[\tau] + K\mathbb{V}(\tau)).$$

*Proof:* Let  $\lambda = \sqrt{\frac{K \log(K) e^2}{T}}$  and note that for  $\Delta > \lambda$ ,  $\log(T\Delta^2)/\Delta$  is decreasing in  $\Delta$ . Then, for constants  $C_1, C_2 > 0$  we can bound the regret in the previous theorem by

$$\mathbb{E}[\mathfrak{R}_T] \leq \sum_{j: \Delta_j \leq \lambda} \mathbb{E}[\mathfrak{R}_t^{(j)}] + \sum_{j: \Delta_j > \lambda} \mathbb{E}[\mathfrak{R}_T^{(j)}] \leq \frac{KC_1 \log(T\lambda^2)}{\lambda} + KC_2(\mathbb{E}[\tau] + \mathbb{V}(\tau)) + T\lambda.$$

substituting in the above value of  $\lambda$  gives a worst case regret bound that scales with  $O(\sqrt{KT \log(K)} + K(\mathbb{E}[\tau] + \mathbb{V}(\tau)))$ .  $\square$

**Remark:** If  $\mathbb{E}[\tau] \geq 1$ , we can replace the  $\mathbb{V}(\tau)$  terms in the regret bounds with  $\mathbb{V}(\tau)/\mathbb{E}[\tau]$ . This follows by using the alternative definition of  $n_m$  suggested in the remark at the end of Section D.1.

## E. Additional Experimental Results

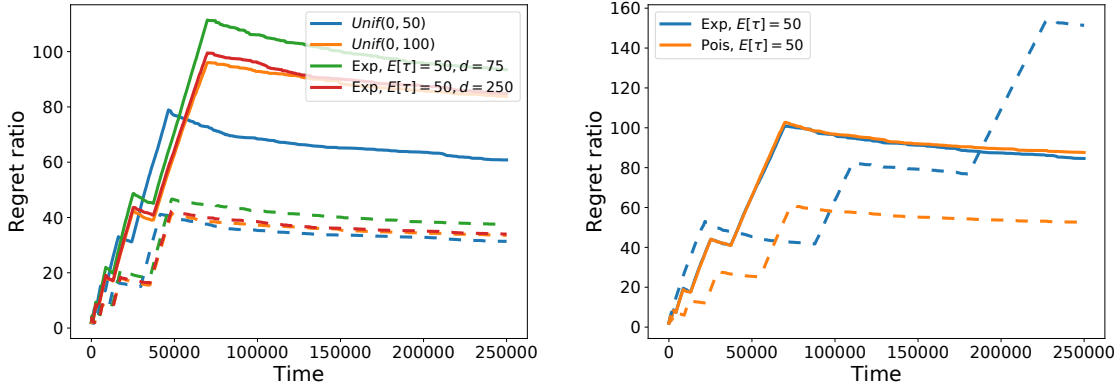
### E.1. Increasing the Expected Delay

Here we investigate the effect of increasing the mean delay on both our algorithm and QPM-D (Joulani et al., 2013) and demonstrate that the regret of both algorithms increases linearly with  $\mathbb{E}[\tau]$ , as indicated by our theoretical results. We use the same experimental set up as described in Section 5. In Figure 5, we are interested in the impact of the mean delay on the regret so we kept the delay distribution family the same, using a  $\mathcal{N}_+(\mu, 100)$  (Normal distribution with mean  $\mu$ , variance 100, truncated at 0) as the delay distribution. We then ran the algorithms for increasing mean delays and plotted the ratio of the regret at  $T$  to the regret of the same algorithm when the delay distribution was  $\mathcal{N}_+(0, 100)$ . In this case, the regret was averaged over 1000 replications for ODAAF and ODAAF-V, and 5000 for QPM-D (this was necessary since the variance of the regret of QPM-D was significant). Here, it can be seen that increasing the mean delay causes the regret of all three algorithms to increase linearly. This is in accordance with the regret bounds which all include a linear factor of  $\mathbb{E}[\tau]$  (since here  $\log(T)$  is kept constant). It can also be seen that ODAAF-V scales better with  $\mathbb{E}[\tau]$  than ODAAF (for constant variance). Particularly, at  $\mathbb{E}[\tau] = 100$ , the relative increase in ODAAF-V is only 1.2 whereas that of ODAAF is 4 (QPM-D has the best relative increase of 1.05).

### E.2. Comparison with Vernade et al. (2017)

Here we compare our algorithms, ODAAF, ODAAF-B and ODAAF-V, to the (non-censored) DUCB algorithm of Vernade et al. (2017). We use the same experimental setup as described in Section 5. As in the comparison to QPM-D, in Figure 6





(a) Bounded delays. Ratios of regret of ODAF (solid lines) and ODAF-B (dotted lines) to that of DUCB.

(b) Unbounded delays. Ratios of regret of ODAF (solid lines) and ODAF-V (dotted lines) to that of DUCB.

Figure 6: The ratios of regret of variants of our algorithm to that of DUCB for different delay distributions.

we plot the ratios of the cumulative regret of our algorithms to that of DUCB for different delay distributions. In Figure 6a, we consider bounded delay distributions and in Figure 6b, we consider unbounded delay distributions. From these plots, we observe that, as in the comparison to QPM-D in Figure 3, the regret ratios all converge to a constant. Thus we can conclude that the order of regret of our algorithms match that of DUCB, even though the DUCB algorithm of Vernade et al. (2017) has considerably more information about the delay distribution. In particular, along with knowledge on the individual rewards of each play (non-anonymous observations), DUCB also uses complete knowledge of the cdf of the delay distribution to re-weight the average reward for each arm. Thus, our algorithms are able to match the rate of regret of Vernade et al. (2017) and QPM-D of Joulani et al. (2013) while just receiving aggregated, anonymous observations and using only knowledge of the expected delay rather than the entire cdf.

We ran the DUCB algorithm with parameter  $\epsilon = 0$ . As pointed out in Vernade et al. (2017), the computational bottleneck in the DUCB algorithm is evaluating the cdf at all past plays of the arms in every round. For bounded delay distributions, this can be avoided using the fact that the cdf will be 1 for plays more than  $d$  steps ago. In the case of unbounded distributions, in order to make our experiments computationally feasible, we used the approximation  $\mathbb{P}(\tau \leq d) = 1$  for  $d \geq 200$ . Another nuance of the DUCB algorithm is due to the fact that in the early stages, the upper confidence bounds are dominated by the uncertainty terms, which themselves involve dividing by the cdf of the delay distributions. The arm that is played last in the initialization period will have the highest cdf and so its confidence bound will be largest and DUCB will play this arm at time  $K + 1$  (and possibly in subsequent rounds unless the cdf increases quickly enough). In order to overcome this, we randomize the order that we play the arms in during the initialization period in each replication of the experiment. Note that we did not run DUCB with half normal delays as DUCB divides by the cdf of the delay distribution and in this case the cdf would be 0 at some points.

## F. Naive Approach for Bounded Delays

In this section we describe a naive approach to defining the confidence intervals when the delay is bounded by some  $d \geq 0$  and show that this leads to sub-optimal regret. Let

$$w_m = \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}} + \frac{md}{n_m}.$$

denote the width of the confidence intervals used in phase  $m$  for any arm  $j$ . We start by showing that the confidence bounds hold with high probability:

**Lemma 31** For any phase  $m$  and arm,  $j$ ,

$$\mathbb{P}(|\bar{X}_{m,j} - \mu_j| > w_m) \leq \frac{2}{T\tilde{\Delta}_m^2}.$$

*Proof:* First note that since the delay is bounded by  $d$ , at most  $d$  rewards from other arms can seep into phase  $i$  of playing arm  $j$  and at most  $d$  rewards from arm  $j$  can be lost. Defining  $S_{i,j}$  and  $U_{i,j}$  as the start and end points of playing arm  $j$  in phase  $i$ , respectively, we have

$$\left| \sum_{t=S_{i,j}}^{U_{i,j}} R_{j,t} - \sum_{t=S_{i,j}}^{U_{i,j}} X_t \right| \leq d, \quad (27)$$

because we can pair up some of the missing and extra rewards, and in each pair the difference is at most one. Then, since  $T_j(m) = \cup_{i=1}^m \{S_{i,j}, S_{i,j} + 1, \dots, U_{i,j}\}$  and using (27) we get

$$\frac{1}{n_m} \left| \sum_{t \in T_j(m)} R_{j,t} - \sum_{t \in T_j(m)} X_t \right| \leq \frac{md}{n_m}.$$

Define  $\bar{R}_{m,j} = \frac{1}{|T_j(m)|} \sum_{t \in T_j(m)} R_{j,t}$  and recall that  $\bar{X}_{m,j} = \frac{1}{|T_j(m)|} \sum_{t \in T_j(m)} X_t$ . For any  $a > \frac{md}{n_m}$ ,

$$\begin{aligned} \mathbb{P}(|\bar{X}_{m,j} - \mu_j| > a) &\leq \mathbb{P}(|\bar{X}_{m,j} - \bar{R}_{m,j}| + |\bar{R}_{m,j} - \mu_j| > a) \leq \mathbb{P}\left(|\bar{R}_{m,j} - \mu_j| > a - \frac{md}{n_m}\right) \\ &\leq 2 \exp \left\{ -2n_m \left( a - \frac{md}{n_m} \right)^2 \right\}, \end{aligned}$$

where the first inequality is from the triangle inequality and the last from Hoeffding's inequality since  $R_{j,t} \in [0, 1]$  are independent samples from  $\nu_j$ , the reward distribution of arm  $j$ . In particular, taking  $a = \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}} + \frac{md}{n_m}$  guarantees that  $\mathbb{P}(|\bar{X}_j - \mu_j| > a) \leq \frac{2}{T\tilde{\Delta}_m^2}$ , finishing the proof.  $\square$

Observe that setting

$$n_m = \left\lceil \frac{1}{2\tilde{\Delta}_m^2} \left( \sqrt{\log(T\tilde{\Delta}_m^2)} + \sqrt{\log(T\tilde{\Delta}_m^2) + 4\tilde{\Delta}_m md} \right)^2 \right\rceil. \quad (28)$$

ensures that  $w_m \leq \frac{\tilde{\Delta}_m}{2}$ . Using this, we can substitute this value of  $n_m$  into Improved UCB and use the analysis from (Auer & Ortner, 2010) to get the following bound on the regret.

**Theorem 32** *Assume there exists a bound  $d \geq 0$  on the delay. Then for all  $\lambda > 0$ , the expected regret of the Improved UCB algorithm run with  $n_m$  defined as in (28) can be upper bounded by*

$$\sum_{\substack{j \in A \\ \Delta_j > \lambda}} \left( \Delta_j + \frac{64 \log(T\Delta_j^2)}{\Delta_j} + 64 \log(2/\Delta_j)d + \frac{96}{\Delta_j} \right) + \sum_{\substack{j \in A \\ 0 < \Delta_j < \lambda}} \frac{64}{\lambda} + T \max_{\substack{j \in A \\ \Delta_j \leq \lambda}} \Delta_j$$

*Proof:* The result follows from the proof of Theorem 3.1 of (Auer & Ortner, 2010) using the above definition of  $n_m$ .  $\square$

In particular, optimizing with respect to  $\lambda$  gives worst case regret of  $O(\sqrt{KT \log K} + Kd \log T)$ . This is a suboptimal dependence on the delay, particularly when  $d \gg \mathbb{E}[\tau]$ .