Double Explore-then-Commit: Asymptotic Optimality and Beyond

Tianyuan Jin TIANYUAN@U.NUS.EDU
School of Computing, National University of Singapore

Pan Xu PANXU@CS.UCLA.EDU
Department of Computer Science, University of California, Los Angeles

Xiaokui Xiao XKXIAO@NUS.EDU.SG
School of Computing, National University of Singapore

Quanquan Gu QGU@CS.UCLA.EDU
Department of Computer Science, University of California, Los Angeles

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

The explore-then-commit (ETC) strategy, which consists of an exploration phase followed by an exploitation phase, is one of the most widely used algorithms in a variety of online decision applications. Nevertheless, it has been shown in Garivier et al. (2016) that ETC is suboptimal in the asymptotic sense as the horizon grows, and thus, is worse than fully sequential strategies such as Upper Confidence Bound (UCB). In this paper, we propose a double explore-then-commit (DETC) algorithm that has two exploration and exploitation phases and show that it can achieve the asymptotically optimal regret bound. To our knowledge, DETC is the first non-fully-sequential algorithm that achieves asymptotic optimality. In addition, we extend DETC to batched bandit problems, where (i) the exploration process is split into a small number of batches and (ii) the round complexity[1] is of central interest. We prove that a batched version of DETC can achieve the asymptotic optimality with only a constant round complexity. This is the first batched bandit algorithm that can attain the optimal asymptotic regret bound and optimal round complexity simultaneously.

Keywords: Multi-armed bandit, regret bound, explore-then-commit, batched bandit, round complexity, asymptotic optimality.

1. Introduction

We study the multi-armed bandit problem, where an agent is asked to choose a bandit arm A_t from a set of arms $\{1, 2, \ldots, K\}$ at every time step t. Then it observes a reward r_t associated with arm A_t following a 1-subgaussian distribution with an unknown mean value μ_{A_t}. For an arbitrary horizon length T, the performance of any strategy for the bandit problem is measured by the *expected cumulative regret*, which is defined as:

$$R_\mu(T) = T \cdot \max_{i\in\{1,2,\cdots,K\}} \mu_i - \mathbb{E}_\mu\big[\textstyle\sum_{t=1}^T r_t\big], \tag{1.1}$$

where the subscript μ denotes the bandit instance consisting of the K arms $\{\mu_1, \ldots, \mu_K\}$.

[1] Round complexity is defined as the total number of times an algorithm needs to update its learning policy. For instance, an UCB algorithm on a bandit problem with time horizon T will have $O(T)$ round complexity because it needs to update its estimation for arms based on the reward collected at each time step.

Assume without loss of generality that arm 1 has the highest mean, i.e., $\mu_1 = \max_i\{\mu_i\}$. Lai and Robbins (1985); Katehakis and Robbins (1995) show that when each arm's reward distribution is Gaussian, the expected regret of any strategy is at least $\sum_{i:\Delta_i>0} 2\log T/\Delta_i$ when T approaches infinity, where $\Delta_i = |\mu_1 - \mu_i|$ denotes the difference between the mean rewards of arm 1 and i. That is,

$$\liminf_{T\to\infty} \frac{R_\mu(T)}{\log T} \geq \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}. \tag{1.2}$$

When Δ_i ($i = 1, 2, \ldots, K$) are known to the decision maker in advance, Garivier et al. (2016) show that the asymptotic lower bound turns to

$$\liminf_{T\to\infty} \frac{R_\mu(T)}{\log T} \geq \sum_{i:\Delta_i>0} \frac{1}{2\Delta_i}. \tag{1.3}$$

We refer to $\lim_{T\to\infty} R_\mu(T)/\log T$ as the *asymptotic regret rate*, and we say that an algorithm is *asymptotically optimal* if it achieves the regret lower bound in (1.2) (when Δ_i are unknown) or (1.3) (when Δ_i are known).

A number of popular multi-armed bandit algorithms such as UCB (Katehakis and Robbins, 1995; Garivier and Cappé, 2011), Thompson Sampling (Agrawal and Goyal, 2017; Korda et al., 2013), and Bayes UCB (Kaufmann et al., 2018) are already asymptotically optimal. However, all of them are *fully sequential*[2] in the sense that they need to observe the outcome of each arm pull before deciding which arm to pull next. Sequential algorithms are unsuitable for applications where each arm pull takes a substantial amount of time. For example, in clinical trials, each treatment involving a human participant can be regarded as an arm pull, and the outcome of the treatment can only be observed after a defined time period. It is thus unaffordable to conduct all treatments in a sequential manner due to the prohibitive total time cost. In the aim of reducing the waiting time for outcomes and taking advantage of parallelism, strategies with distinct exploration and exploitation stages are often more preferable compared with fully-sequential algorithms, where a batch of arm pulls are conducted simultaneously within any stage and the outcomes of the entire batch are only needed when a stopping criteria for stage switching is satisfied.

The most natural approach for separating the exploration and exploitation stages is to first uniformly pull each arm for the same number of times (*the exploration stage*), and then pull the arm with the larger average reward repeatedly based on the result in the previous stage (*the exploitation stage*). Such strategies with distinct exploration and exploitation stages fall into the class of approaches named explore-then-commit (ETC) (Perchet et al., 2016; Garivier et al., 2016), which are simple and widely implemented in various online applications, such as clinical trials, crowdsourcing and marketing (Perchet et al., 2016; Garivier et al., 2016; Gao et al., 2019). When the length of the exploration stage (or equivalently the batch size of the pulls for each arm) is a fixed number, such an strategy is referred to as FB-ETC (Garivier et al., 2016). In this case we only need to collect the outcomes of the arm pulls at the end of the exploration stage, which is a one-time-only effort.

Despite the significant improvement in efficiency and parallelism compared with fully sequential algorithms, the regret of such two-stage ETC strategies is shown to be near-optimal with careful tuning (Garivier and Kaufmann, 2016) but is also essentially suboptimal (Garivier et al., 2016) in

[2] In particular, we say a strategy is fully sequential if its round complexity is polynomial in the horizon length T. Similarly, we say a strategy is non-fully-sequential if its round complexity is $o(T)$.

the sense that they cannot achieve the exact asymptotically optimal lower bounds in (1.2) or (1.3). Following these nice and clean theoretical results, a natural and open question is:

Can non-fully-sequential strategies such as ETC strategies with multiple stages achieve the asymptotically optimal regret?

In this paper, we answer this question affirmatively by proposing a double explore-then-commit (DETC) algorithm that consists of two exploration and two exploitation stages, which directly improves the ETC algorithm proposed in Garivier et al. (2016). Take the two-armed bandit problem as an example, the key idea of DETC is illustrated as follows: based on the result of the first exploration stage, the algorithm will commit to the arm with the largest average reward and pull it for a long time in the exploitation stage. After the first exploitation stage, the algorithm will have a confident estimate of the chosen arm. However, since the unchosen arm is never pulled after the first exploration stage, the algorithm is still not sure whether the unchosen arm is underestimated. Therefore, a second exploration stage for the algorithm to pull the unchosen arms is necessary. After this stage, the algorithm will have sufficiently accurate estimate for all arms and just needs another exploitation stage to commit to the arm with the largest average reward. In contrast to the above double explore-then-commit algorithm, existing ETC algorithms may have inaccurate estimates for both the optimal arm and the suboptimal arms and hence suffers a suboptimal regret.

1.1. Our Contributions

Double explore-then-commit with asymptotic optimality We first study two-armed bandits, where we simplify the notation by writing $\Delta = \Delta_2 = |\mu_1 - \mu_2|$ as the gap. When Δ is a known parameter to the algorithm, we prove DETC (Algorithm 1) achieves the asymptotically optimal regret rate $1/(2\Delta)$, the instance-dependent optimal regret $O(\log(T\Delta^2)/\Delta)$ and the minimax optimal regret $O(\sqrt{T})$ for two-armed bandits. Our result significantly improves the $4/\Delta$ asymptotic regret rate of FB-ETC with a fixed exploration length and the $1/\Delta$ asymptotic regret rate of SPRT-ETC with a data-dependent exploration length (Garivier et al., 2016).

When Δ is unknown, we prove that DETC (Algorithm 2) achieves the asymptotically optimal regret rate $2/\Delta$, the instance-dependent regret $O(\log(T\Delta^2)/\Delta)$ and the minimax optimal regret $O(\sqrt{T})$. This again improves the $4/\Delta$ asymptotic regret rate of the BAI-ETC algorithm proposed in Garivier et al. (2016). In both the known gap and the unknown gap settings, this is the first time that the regrets of ETC algorithms have been proved to match the asymptotic lower bounds. In contrast, Garivier et al. (2016) proved that the $1/\Delta$ asymptotic regret rate for the known gap case and the $4/\Delta$ asymptotic regret rate for the unknown gap case are not improvable for the two-stage ETC algorithms, which justifies the essence of the double exploration technique used in DETC in order to achieve the asymptotic regret.

We also propose a variant of DETC (Algorithm 3) with unknown gaps that is simultaneously instance-dependent/minimax optimal and asymptotically optimal for two-armed bandits. Our analysis and algorithmic framework also suggests an effective way of combining the asymptotically optimal DETC algorithm with any other minimax optimal algorithms to achieve the instance-dependent/minimax and asymptotic optimality simultaneously. To promote the application of DETC in practice when the time horizon T is unknown ahead of time, we also proposed an anytime version of DETC (Algorithm 6) using the doubling trick which enjoys the asymptotic optimal regret. Moreover, we extend our DETC algorithm to K-armed bandits and prove that DETC (Algorithm

7) achieves the asymptotically optimal regret rate $\sum_{i:\Delta_i>0} 2/\Delta_i$ for K-armed bandits (Lai and Robbins, 1985), where Δ_i is the gap between the best arm and the i-th arm, $i \in [K]$.

Double explore-then-commit in batched bandits The most direct application of explore-then-commit strategies is its perfect fit into the *batched bandit* problem (Perchet et al., 2016; Bertsimas and Mersereau, 2007; Chick and Gans, 2009; Agarwal et al., 2017; Gao et al., 2019; Esfandiari et al., 2019), which requires arms to be pulled in *rounds*. In each round, we are allowed to pull multiple arms at the same time, but can only observe the outcomes at the end of the round. The central question is to minimize not only the expected cumulative regret after T arm pulls, but also the number of rounds (*round complexity*) that we check the outcome of the pulls.

We prove that a simple variant of DETC (Algorithm 4 and Algorithm 5 for the known gap and unknown gap settings respectively) can achieve $O(1)$ round complexity while maintaining the asymptotically optimal regret for two-armed bandits. This is a significant improvement of the round complexity of fully sequential strategies such as UCB and UCB2 (Lai and Robbins, 1985; Auer et al., 2002a; Garivier and Cappé, 2011), which usually requires $O(T)$ or $O(\log T)$ rounds. Our result suggests that it is not necessary to use the outcome at each time step as in fully sequential algorithms such as UCB to achieve the asymptotic optimality. Existing batched bandit algorithms (Perchet et al., 2016; Gao et al., 2019; Esfandiari et al., 2019) are based on two-stage ETC, and hence is suboptimal in the asymptotic sense. In contrast, DETC is the first batched bandit algorithm that achieves the asymptotic optimality in regret and the optimal round complexity.

Notation We denote $\log^+(x) = \max\{0, \log x\}$. We use notations $\lfloor x \rfloor$ (or $\lceil x \rceil$) to denote the largest integer that is no larger (or no smaller) than x. For any functions f and g, we use $f(T) = O(g(T))$ to imply that $f(T) \le Cg(T)$ for some constant $C > 0$ that is independent of T. We use $f(T) = o(g(T))$ to imply that $\lim_{T\to\infty} f(T)/g(T) = 0$. A random variable X is said to follow 1-subgaussian distribution, if it holds that $\mathbb{E}_X[\exp(\lambda X - \lambda\mathbb{E}_X[X])] \le \exp(\lambda^2/2)$ for all $\lambda \in \mathbb{R}$.

2. Double Explore-then-Commit Strategies

The vanilla ETC strategy (Perchet et al., 2016; Garivier et al., 2016) consists of two stages: in stage one (the exploration stage), the agent pulls all arms for the same number of times, which can be a fixed integer or a data-dependent stopping time, leading to the FB-ETC and SPRT-ETC (or BAI-ETC) algorithms in (Garivier et al., 2016); in stage two (the exploitation stage), the agent pulls the arm that achieves the best average reward according to the outcome of stage one. As we mentioned in the introduction, none of these algorithms can achieve the asymptotic optimality in (1.2) or (1.3). To tackle this problem, we propose a double explore-then-commit strategy for two-armed bandits that improves ETC to be asymptotically optimal while still keeping non-fully-sequential.

2.1. Warm-Up: The Known Gap Setting

We first consider the case where the gap $\Delta = \mu_1 - \mu_2$ is known to the decision maker (recall that we assume w.l.o.g. that arm 1 is the optimal arm). We propose a double explore-then-commit (DETC) algorithm, which consists of four stages. The details are displayed in Algorithm 1.

At the initialization step, we pull both arms once, after which we set the current time step $t = 2$. In *Stage I*, DETC pulls both arms for $\tau_1 = 4\lceil\log(T_1\Delta^2)/\Delta^2\rceil$ times respectively, where both τ_1 and T_1 are predefined parameters. At time step t, we define $T_k(t)$ to be the total number of times that arm k ($k = 1, 2$) has been pulled so far, i.e., $T_k(t) = \sum_{i=1}^{t} \mathbb{1}_{\{A_i=k\}}$, where A_i

Algorithm 1 Double Explore-then-Commit (DETC) in the Known Gap Setting

```
input $T$, $\epsilon_T$ and $\Delta$
Initialization: Pull arms $A_1 = 1$ and $A_2 = 2$, $t \leftarrow 2$, $T_1 = \lceil 2\log(T\Delta^2)/(\epsilon_T^2 \cdot \Delta^2)\rceil$, $\tau_1 = 4\lceil \log(T_1\Delta^2)/\Delta^2\rceil$
Stage I: Explore all arms uniformly
while $t \le 2\tau_1$ do
    Pull arms $A_{t+1} = 1$ and $A_{t+2} = 2$, $t \leftarrow t + 2$
end while
Stage II: Commit to the arm with the largest average reward
$1' \leftarrow \arg\max_{k\in\{1,2\}} \widehat{\mu}_k(t)$
while $T_{1'}(t) \le T_1$ do
    Pull arm $A_{t+1} = 1'$, $t \leftarrow t + 1$
end while
Stage III: Explore the unchosen arm in Stage II
$\mu' \leftarrow \widehat{\mu}_{1'}(t)$, $t_2 \leftarrow 0$, $2' \leftarrow \{1,2\} \setminus 1'$, $\theta_{2',0} \leftarrow 0$
while $2(1-\epsilon_T)t_2\Delta \mid \mu' - \theta_{2',t_2} \mid < \log(T\Delta^2)$ do
    Pull arm $A_{t+1} = 2'$ and observe reward $r_{t+1}$
    $\theta_{2',t_2+1} = (t_2\theta_{2',t_2} + r_{t+1})/(t_2+1)$, $t \leftarrow t+1$, $t_2 \leftarrow t_2 + 1$
end while
Stage IV: Commit to the arm with the largest average reward
$a \leftarrow 1'\,\mathbb{1}\{\widehat{\mu}_{1'}(t) \ge \theta_{2',t_2}\} + 2'\,\mathbb{1}\{\widehat{\mu}_{1'}(t) < \theta_{2',t_2}\}$
while $t \le T$ do
    Pull arm $a$, $t \leftarrow t + 1$
end while
```

is the arm pulled at time step i. Then we define the average reward of arm k at time step t as $\widehat{\mu}_k(t) := \sum_{i=1}^{t} \mathbb{1}_{\{A_i = k\}}\, r_i / T_k(t)$, where r_i is the reward received by the algorithm at time i.

In *Stage II*, DETC repeatedly pulls the arm with the largest average reward at the end of *Stage I*, denoted by arm $1' = \arg\max_{k=1,2} \widehat{\mu}_{k,\tau_1}$, where $\widehat{\mu}_{k,\tau_1}$ is the average reward of arm k after its τ_1-*th* pull. Note that before *Stage II*, arm $1'$ has been pulled for τ_1 times. We will terminate *Stage II* after the total number of pulls of arm $1'$ reaches T_1. It is worth noting that *Stage I* and *Stage II* are similar to existing ETC algorithms (Garivier et al., 2016), where these two stages are referred to as the *Explore* (explore different arms) and the *Commit* (commit to one single arm) stages respectively.

The key difference here is that instead of pulling arm $1'$ till the end of the horizon (time step T), our Algorithm 1 sets a check point $T_1 < T$. After arm $1'$ has been pulled for T_1 times, we stop and check the average reward of the arm that is not chosen in *Stage II*, denoted by arm $2'$. The motivation for this halting follows from a natural question: *What if we have committed to the wrong arm?* Even though arm $2'$ is not chosen based on the outcome of *Stage I*, it can still be optimal due to random sampling errors. To avoid such a case, we pull arm $2'$ for more steps such that the average rewards of both arms can be distinguished from each other. Specifically, in *Stage III* of Algorithm 1, arm $2'$ is repeatedly pulled until

$$2(1-\epsilon_T)t_2\Delta|\mu' - \theta_{2',t_2}| \ge \log(T\Delta^2), \tag{2.1}$$

where $\epsilon_T > 0$ is a parameter, t_2 is the recalculated number of pulls in Stage *III* at time step t[3], θ_{2',t_2} is the average reward of arm $2'$ in *Stage III* and μ' is the average reward of arm $1'$ recorded at the end of *Stage II*. Note that $\mu' = \widehat{\mu}_{1'}(t)$ throughout *Stage III* since arm $1'$ is not pulled in this stage.

As is discussed in the above paragraph, at the end of *Stage II*, the average reward μ' for arm $1'$ already concentrates on its expected reward. Therefore, in *Stage III* of DETC, the sampling error only comes from pulling arm $2'$. Hence, our DETC algorithm offsets the drawback ETC algorithms where the sampling error comes from both arms. In the remainder of the algorithm (*Stage IV*), we just again commit to the arm with the largest empirical reward from at the end of *Stage III*.

Now, we present the regret bound of Algorithm 1. Note that if $T\Delta^2 < 1$, the worst case regret is trivially bounded by $T\Delta < \sqrt{T}$ and the asymptotic regret rate is meaningless since $\Delta \to 0$ when $T \to \infty$. Hence, in the following theorem, we assume $T\Delta^2 \geq 1$.

Theorem 2.1 *If ϵ_T is chosen such that $T_1\Delta^2 \geq 1$, the regret of Algorithm 1 is upper bounded as*

$$R_\mu(T) \leq 2\Delta + \frac{8}{\Delta} + \frac{4\log(T_1\Delta^2)}{\Delta} + \frac{\log(T\Delta^2)}{2(1-\epsilon_T)^2\Delta} + \frac{2\sqrt{\log(T\Delta^2)}+2}{(1-\epsilon_T)^2\Delta}. \tag{2.2}$$

In particular, let $\epsilon_T = \min\{\sqrt{\log(T\Delta^2)/(\Delta^2\log^2 T)}, 1/2\}$, then $\limsup_{T\to\infty} R_\mu(T)/\log T \leq 1/(2\Delta)$, and $R_\mu(T) = O(\Delta + \log(T\Delta^2)/\Delta)$.

The proof of Theorem 2.1 can be found in Section C.1. This theorem states that Algorithm 1 achieves the asymptotically optimal regret rate $1/(2\Delta)$ and instance-dependent optimal regret $O(\Delta + 1/\Delta\log(T\Delta^2))$, when parameter ϵ_T is properly chosen. In comparison, the ETC algorithm in Garivier et al. (2016) can only achieve $1/\Delta$ asymptotic regret rate under the same setting, which is suboptimal for multi-armed bandit problems (Lai and Robbins, 1985) when gap Δ is known to the decision maker. It is important to note that, Garivier et al. (2016) also proved a lower bound for asymptotic optimality of ETC and showed that the $1/\Delta$ asymptotic regret rate of 'single' explore-then-commit algorithms cannot be improved. Therefore, the double exploration techniques in our DETC is indeed essential for breaking the $1/\Delta$ barrier in the asymptotic regret rate.

The asymptotic optimality is also achieved by the Δ-UCB algorithm in Garivier et al. (2016), which is a fully sequential strategy. In stark contrast, DETC shows that non-fully-sequential algorithms can also achieve the asymptotically optimal regret for multi-armed bandit problems. Compared with Δ-UCB, DETC has distinct stages of exploration and exploitation which makes the implementation simple and more practical. A more important and unique feature of DETC is its lower round complexity for batched bandit problems, which will be thoroughly discussed in Section 3.1.

Note that Algorithm 1 is a non-fully-sequential strategy in the sense that it only needs to update the policy for $o(T)$ times. In particular, let us take a close look at the round complexity of Algorithm 1. Since we do not need the outcomes of arm pulls within *Stages I, II* and *IV*, we only need to collect the rewards and update the policy at the end of these stages. In other words, *Stages I, II* and *IV* only need three rounds in total. Besides, according to Theorem 2.1, the expected number of pulls of *Stage III* is $\log T/(2\Delta^2)$. Therefore, the expected number of policy updates in Algorithm 1 is $O(\log T)$, which is much smaller than $O(T)$, the round complexity of fully sequential strategies. In Section 3.1, we will show that by carefully choosing the batch sizes used in *Stage III*, we can modify Algorithm 1 and further improve the round complexity from $O(\log T)$ to $O(1)$.

[3] For the simplicity of analysis, we recalculate the average reward of arm $2'$ in *Stage III*. However, it is possible to re-use the arms pulls in *Stage I* and *II*, without any major changes in the conclusions (see Jin et al. (2021) for example).

Algorithm 2 Double Explore-then-Commit (DETC) in the Unknown Gap Setting

```
input $T, T_1$
Initialization: Pull arms $A_1 = 1$, $A_2 = 2$, $t \leftarrow 2$;
Stage I: Explore all arms uniformly
while $| \widehat{\mu}_1(t) - \widehat{\mu}_2(t) | < \sqrt{16/t \log^+(T_1/t)}$ do
    Pull arms $A_{t+1} = 1$ and $A_{t+2} = 2$, $t \leftarrow t + 2$
end while
Stage II: Commit to the arm with the largest average reward
$1' \leftarrow \arg\max_i \widehat{\mu}_i(t)$
while $T_{1'}(t) \le T_1$ do
    Pull arm $A_{t+1} = 1'$, $t \leftarrow t + 1$
end while
Stage III: Explore the unchosen arm in Stage II
$\mu' \leftarrow \widehat{\mu}_{1'}(t)$, $2' \leftarrow \{1, 2\} \setminus 1'$
Pull arm $A_{t+1} = 2'$ and observe reward $r_{t+1}$, $\theta_{2',1} = r_{t+1}$, $t \leftarrow t + 1$, $t_2 \leftarrow 1$
while $|\mu' - \theta_{2',t_2}| < \sqrt{2/t_2 \log\big(T/t_2\big(\log^2(T/t_2) + 1\big)\big)}$ do
    Pull arm $A_{t+1} = 2'$ and observe reward $r_{t+1}$
    $\theta_{2',t_2+1} = (t_2\theta_{2',t_2} + r_{t+1})/(t_2 + 1)$, $t \leftarrow t + 1$, $t_2 \leftarrow t_2 + 1$
end while
Stage IV: Commit to the arm with the largest average reward
$a \leftarrow 1' \, \mathbb{1}\{\widehat{\mu}_{1'}(t) \ge \theta_{2',t_2}\} + 2' \, \mathbb{1}\{\widehat{\mu}_{1'}(t) < \theta_{2',t_2}\}$
while $t \le T$ do
    Pull arm $a$, $t \leftarrow t + 1$
end while
```

2.2. Double Explore-then-Commit in the Unknown Gap Setting

In real world applications, the gap Δ is often unknown. Thus, it is favorable to design an algorithm without the knowledge of Δ. However, this imposes issues with Algorithm 1, since the stopping rules of the two exploration stages (*Stage I* and *Stage III*) are unknown. To address this challenge, we propose a DETC algorithm where the gap Δ is unknown to the decision maker, which is displayed in Algorithm 2.

Similar to Algorithm 1, Algorithm 2 also consists of four stages, where *Stage I* and *Stage III* are double exploration stages that ensure we have chosen the right arm to pull in the subsequent stages. Since we have no knowledge about Δ, we derive the stopping rule for *Stage I* by comparing the empirical average rewards of both arms. Once we have obtained empirical estimates of the mean rewards that are able to distinguish two arms in the sense that $|\widehat{\mu}_1(t) - \widehat{\mu}_2(t)| \ge \sqrt{16 \log^+(T_1/t)/t}$, we terminate *Stage I*. Here t is the current time step of the algorithm and T_1 is a predefined parameter. Similar to Algorithm 1, based on the outcomes of *Stage I*, we commit to arm $1' = \mathrm{argmax}_{i=1,2} \widehat{\mu}_i(t)$ at the end of *Stage I* and pull this arm repeatedly throughout *Stage II*. In *Stage III*, we turn to pull arm $2'$ that is not chosen in *Stage II* until the average reward of arm $2'$ is significantly larger or smaller than that of arm $1'$ chosen in *Stage II*. In *Stage IV*, we again commit to the best empirically preforming arm and pull it till the end of the algorithm.

Compared with Algorithm 1, in both exploration stages of Algorithm 2, we do not use the information of the gap Δ at the cost of sequentially deciding the stopping rule in these two stages. In the following theorem, we present the regret bound of Algorithm 2 and show that this regret is still asymptotically optimal.

Theorem 2.2 *Let $T_1 = \log^2 T$, then the regret of Algorithm 2 satisfies*

$$\lim_{T\to\infty} R_\mu(T)/\log T = 2/\Delta.$$

The proof of Theorem 2.2 can be found in Section C.2. Here we provide some comparison between existing algorithms and Algorithm 2. For two-armed bandits, Lai and Robbins (1985) proved that the asymptotically optimal regret rate is $2/\Delta$. This optimal bound has been achieved by a series of fully sequential bandit algorithms such as UCB (Garivier and Cappé, 2011; Lattimore, 2018), Thompson sampling (Agrawal and Goyal, 2017), Ada-UCB (Kaufmann et al., 2018), etc. All these algorithms are fully sequential, which means they have to examine the outcome from current pull before it can decide which arm to pull in the next time step. In contrast, DETC (Algorithm 2) is non-fully-sequential and separates the exploration and exploitation stages, which is much more practical in many real world applications such as clinical trials and crowdsourcing. In particular, DETC can be easily adapted to batched bandits and achieve a much smaller round complexity than these fully sequential algorithms. We will elaborate this in Section 3.2.

Compared with other ETC algorithms in the unknown gap setting, Garivier et al. (2016) proved a lower bound $4/\Delta$ for 'single' explore-then-commit algorithms, while the regret upper bound of DETC is improved to $2/\Delta$. Therefore, in order to break the $4/\Delta$ barrier in the asymptotic regret rate, our double exploration technique in Algorithm 2 is crucial. Different from DETC in the known gap setting, Theorem 2.2 does not say anything about the minimax or instance-dependent optimality of Algorithm 2. Because we need to guess the gap Δ during the exploration process in *Stage I* and *Stage III*, additional errors may be introduced if the guess is not accurate enough. We will discuss this in details in the next section.

2.3. Minimax and Asymptotically Optimal DETC

If we compare the stopping rules of the exploration stages in Algorithm 1 and Algorithm 2, we can observe that the stopping rule in the known gap setting (Algorithm 1) depends on the gap Δ (more specifically, it depends on the quantity $1/\Delta^2$ according to our analysis of the theorems in the appendix). In Algorithm 2, the gap Δ is unknown and guessed by the decision maker. This causes problems when the unknown Δ is too small (e.g., $\Delta = 1/T^{0.1}$), where $1/\Delta^2$ is significantly large than $\log^2 T$. In this case, after $T_1 = \log^2 T$ pulls of arm $1'$ in *Stage II* of Algorithm 2, the average reward of $1'$ may not be close to its mean reward within a Δ range. Hence, it fails to achieve the instance-dependent/minimax optimality.

Now we are going to show that a simple variant of Algorithm 2 with additional stopping rules is simultaneously minimax/instance-dependent order-optimal and asymptotically optimal.

The new algorithm is displayed in Algorithm 3 which has the same input, initialization, *Stage I* and *Stage II* as Algorithm 2. In *Stage III*, we add an additional stopping rule $t_2 < \log^2 T$ and everything else remains unchanged as in Algorithm 2. The most notable change is in *Stage IV* of Algorithm 3. Instead of directly committing to the arm with the largest average reward, we will first find out how many pulls are required in *Stage III* to distinguish the two arms. The number of pulls

Algorithm 3 Minimax and Asymptotically Optimal DETC in the Unknown Gap Setting

```
input $T$, $T_1$
Initialization: Pull arms $A_1 = 1$, $A_2 = 2$, $t \leftarrow 2$;
Stage I: Explore all arms uniformly                          (same as in Algorithm 2)
Stage II: Commit to the arm with the largest average reward  (same as in Algorithm 2)
Stage III: Explore the unchosen arm in Stage II
$\mu' \leftarrow \widehat{\mu}_{1'}(t)$, $2' \leftarrow \{1, 2\} \setminus 1'$
Pull arm $A_{t+1} = 2'$ and observe reward $r_{t+1}$, $\theta_{2',1} = r_{t+1}$, $t \leftarrow t + 1$, $t_2 \leftarrow 1$
while $|\mu' - \theta_{2',t_2}| < \sqrt{2/t_2 \log\big(eT/t_2\big(\log^2(T/t_2) + 1\big)\big)}$ and $t_2 < \log^2 T$ do
    Pull arm $A_{t+1} = 2'$ and observe reward $r_{t+1}$
    $\theta_{2',t_2+1} = (t_2\theta_{2',t_2} + r_{t+1})/(t_2 + 1)$, $t \leftarrow t + 1$, $t_2 \leftarrow t_2 + 1$
end while
Stage IV: Commit to the arm with the largest average reward
if $t_2 < \log^2 T$ then
    $a \leftarrow 1' \, \mathbb{1}\{\widehat{\mu}_{1'}(t) \geq \theta_{2',t_2}\} + 2' \, \mathbb{1}\{\widehat{\mu}_{1'}(t) < \theta_{2',t_2}\}$
    while $t \leq T$ do
        Pull arm $a$, $t \leftarrow t + 1$
    end while
else
    Pull arms $A_{t+1} = 1$, $A_{t+2} = 2$ and observe rewards $r_{t+1}$ and $r_{t+2}$
    $p_{1,1} = r_{t+1}$, $p_{2,1} = r_{t+2}$, $t \leftarrow t + 2$, $s \leftarrow 1$
    while $|p_{1,s} - p_{2,s}| < \sqrt{8/s \log^+(T/s)}$ do
        Pull arms $A_{t+1} = 1$ and $A_{t+2} = 2$, and observe rewards $r_{t+1}$ and $r_{t+2}$
        $p_{1,s+1} = (s \cdot p_{1,s} + r_{t+1})/(s + 1)$, $p_{2,s+1} = (s \cdot p_{2,s} + r_{t+2})/(s + 1)$
        $t \leftarrow t + 2$, $s \leftarrow s + 1$
    end while
    $a \leftarrow 1 \, \mathbb{1}\{p_{1,s} \geq p_{2,s}\} + 2 \, \mathbb{1}\{p_{2,s} \geq p_{1,s}\}$
    while $t \leq T$ do
        Pull arm $a$, $t \leftarrow t + 1$.
    end while
end if
```

in *Stage III* is denoted by t_2. If $t_2 < \log^2 T$, then we just commit to the arm with the largest average reward and pull it till the end of the algorithm. This will be the same as in Algorithm 2.

However, if $t_2 = \log^2 T$, this would mean that the gap Δ between two arms is extremely small. In fact, we will prove that if we need to pull arm $2'$ for $\log^2 T$ times to distinguish it from arm $1'$, then with high probability the gap Δ is very small. Consequently, we need to explore both arms again to obtain accurate estimate of their mean rewards. In a nutshell, the early stopping rule $t_2 \geq \log^2 T$ helps us detect the scenario with small Δ with high probability, which ensures the minimax/instance-dependent optimality of Algorithm 3. Moreover, we will show that this condition is only violated with a tiny probability that goes to zero as $T \to \infty$, which ensures that the regret is still asymptotically optimal.

Theorem 2.3 *Let $T_1 = \log^{10} T$. Assume $T\Delta^2 \geq 16e^3$, then the regret of Algorithm 3 satisfies*

$$\lim_{T\to\infty} R_\mu(T)/\log T = 2/\Delta \qquad \textit{and} \qquad R_\mu(T) = O(\Delta + \log(T\Delta^2)/\Delta) = O(\Delta + \sqrt{T}).$$

The proof of Theorem 2.3 can be found in Section C.3. This theorem states that Algorithm 3 achieves the instance-dependent/minimax and the asymptotic optimality regret simultaneously. This is the first ETC-type algorithm that achieves these three optimal regrets simultaneously. Compared with Algorithm 2 presented in the previous subsection, Algorithm 3 has a more complicated implementation of *Stage IV* in the sense that it may have to start over from the uniform exploration (Lines 16-19) and then commit to the arm that it believes to be the optimal arm. Though we prove that the aforementioned event only happens with a small probability (in Section C.3) and we show that Algorithm 3 indeed achieves a better regret bound, it essentially needs more rounds than Algorithm 2 since the *Stage IV* of Algorithm 3 is fully sequential from Line 16 to Line 19.

Apart from the advantages of achieving these optimalities at the same time, we emphasize that Algorithm 3 also provides a framework on how to combine an asymptotically optimal algorithm with a minimax/instance-dependent optimal algorithm. Specifically, in Algorithm 3, the first part (Lines 1-11) of Algorithm 3 ensures the asymptotic optimality and the second part (Lines 14-23) of Algorithm 3 ensures the minimax/instance-dependent optimality. Following our proof of Theorem 2.3 in Section C.3, one can easily verify that the second part (Lines 14-23) can be replaced by any other algorithm that is instance dependent optimal and Theorem 2.3 still holds. The main reason that two optimality algorithm can be combined here is that: (i): the asymptotic optimality focuses on the case that $T \to \infty$, and hence T should dominate $1/\Delta$; (ii) the minimax optimality focuses on the worst case bandits for a fixed T, and hence Δ could be very small (e.g., $\Delta = 1/T^{0.1}$); (iii) our framework can detect if Δ is very small via the stopping rule $t_2 < \log^2 T$ in Line 4 of Algorithm 3.

3. Asymptotically Optimal DETC in Batched Bandit Problems

The proposed DETC algorithms in this paper can be easily extended to batched bandit problems (Perchet et al., 2016; Gao et al., 2019). In this section, we present simple modifications to Algorithms 1 and 2 which we refer to as *Batched DETC*. We prove that they not only achieve the asymptotically optimal regret bounds but also enjoy $O(1)$ round complexities.

3.1. Batched DETC in the Known Gap Setting

We use the same notations that are used in Section 2.1. The Batched DETC algorithm is identical to Algorithm 1 except the stopping rule of *Stage III*. More specifically, let $\tau_0 = \log(T\Delta^2)/(2(1-\epsilon_T)^2\Delta^2)$. In *Stage III* of Algorithm 1, instead of querying the result θ_{2',t_2} at every step $t_2 = 0, 1, \ldots$, we only query it at the following time grid:

$$\mathcal{T} = \left\{ \left\lceil \tau_0 + \frac{2\sqrt{\log(T\Delta^2)}+4}{2(1-\epsilon_T)^2\Delta^2} \right\rceil, \left\lceil \tau_0 + \frac{2(2\sqrt{\log(T\Delta^2)}+4)}{2(1-\epsilon_T)^2\Delta^2} \right\rceil, \left\lceil \tau_0 + \frac{3(2\sqrt{\log(T\Delta^2)}+4)}{2(1-\epsilon_T)^2\Delta^2} \right\rceil, \cdots \right\}. \quad (3.1)$$

At each time point $t_2 \in \mathcal{T}$, we query the results of the bandits pulled since the last time point. Between two time points, we pull the arm $2'$ without accessing the results. The period between two times points is also referred to as a round (Perchet et al., 2016). Reducing the total number of queries, namely, the round complexity, is an important research topic in the batched bandit problem. For the convenience of readers, we present the Batched DETC algorithm for known gaps in Algorithm 4. Note that in this batched version, *Stage I*, *Stage II* and *Stage IV* of Algorithm 4 are identical to that of Algorithm 1, and we omit them for the simplicity of presentation.

Now, we present the round complexity of the Batched DETC in Algorithm 4.

Algorithm 4 Batched DETC in the Known Gap Setting

```
input $T$, $\epsilon_T$, $\Delta$ and $\mathcal{T}$ defined in (3.1)
Initialization: Pull arms $A_1 = 1$ and $A_2 = 2$, $t \leftarrow 2$, $T_1 = \lceil 2\log(T\Delta^2)/(\epsilon_T^2 \cdot \Delta^2) \rceil$, $\tau_1 = 4\lceil \log(T_1\Delta^2)/\Delta^2 \rceil$
Stage I: Explore all arms uniformly                              (same as in Algorithm 1)
Stage II: Commit to the arm with the largest average reward      (same as in Algorithm 1)
Stage III: Explore the unchosen arm in Stage II
$\mu' \leftarrow \widehat{\mu}_{1'}(t)$, $t_2 \leftarrow 0$, $2' \leftarrow \{1,2\} \setminus 1'$, $\theta_{2',s}$ is the recalculated average reward of arm $2'$ after its $s$-th pull in Stage III and $\theta_{2's} \leftarrow 0$, for $s = 0$
while true do
  if $t_2 \in \mathcal{T}$ then
    if $2(1-\epsilon_T)t_2\Delta \mid \mu' - \theta_{2',t_2} \mid \geq \log(T\Delta^2)$ then
      break
    end if
  end if
  Pull arm $A_{t+1} = 2'$, $t \leftarrow t+1$, $t_2 \leftarrow t_2 + 1$
end while
Stage IV: Commit to the arm with the largest average reward      (same as in Algorithm 1)
```

Theorem 3.1 *In the batched bandit problem, the expected number of rounds used in Algorithm 4 is $O(1)$. At the same time, the regret of Algorithm 4 is asymptotically optimal.*

Remark 3.2 *The proof of Theorem 3.1 can be found in Section D.1. Compared with fully sequentially adaptive bandit algorithms such as UCB, which needs $O(T)$ rounds of queries, our DETC algorithm only needs constant rounds of queries (independent of the horizon length T). Compared with another constant round algorithm FB-ETC in (Garivier et al., 2016), our DETC algorithm simultaneously improves the asymptotic regret rate of FB-ETC (i.e., $4/\Delta$) by a factor of* 8.

3.2. Batched DETC in the Unknown Gap Setting

In the unknown gap setting, both the stopping rules of *Stage I* and *Stage III* in Algorithm 2 need to be modified. In what follows, we describe a variant of Algorithm 2 that only needs to check the outcomes at certain time points in *Stage I* and *Stage III*. In particular, let $T_1 = \log^2 T$. In *Stage I*, we query the results and test the condition in Line 3 of Algorithm 5 at the following time grid:

$$t \in \mathcal{T}_2 = \{2\sqrt{\log T}, 4\sqrt{\log T}, 6\sqrt{\log T}, \ldots\}. \tag{3.2}$$

In *Stage III*, the condition in Line 10 of Algorithm 5 is only checked at the following time grid.

$$\begin{aligned} t_2 \in \mathcal{T}_2' = \big\{N_1, & 2/\widehat{\Delta}^2 N_2 \log(T\log^3 T) + 1/\widehat{\Delta}^2 N_2 (\log T)^{\frac{2}{3}}, \\ & 2/\widehat{\Delta}^2 N_2 \log(T\log^3 T) + 2/\widehat{\Delta}^2 N_2 (\log T)^{\frac{2}{3}}, \\ & 2/\widehat{\Delta}^2 N_2 \log(T\log^3 T) + 3/\widehat{\Delta}^2 N_2 (\log T)^{\frac{2}{3}}, \cdots, \log^2 T\big\}. \end{aligned} \tag{3.3}$$

where $N_1 = (2\log T)/\log\log T$, $N_2 = (1 + (\log T)^{-\frac{1}{4}})^2$, and $\widehat{\Delta} = |\mu' - \theta_{2',N_1}|$ is an estimate of Δ' based on the test result after the first round (the first N_1 steps). Apart from restricting $t_2 \in \mathcal{T}_2'$, another difference here from Algorithm 2 is that we require $t_2 \leq \log^2 T$. Thus we will terminate *Stage III* after at most $\log^2 T$ pulls of arm $2'$. For the convenience of readers, we display the modified Algorithm 2 for batched bandits with unknown gaps in Algorithm 5.

Algorithm 5 Batched DETC in the Unknown Gap Setting

```
input $T$, $T_1$, $\mathcal{T}_2$ defined in (3.2), and $\mathcal{T}_2'$ defined in (3.3)
Initialization: Pull arms $A_1 = 1$, $A_2 = 2$, $t \leftarrow 2$
Stage I: Explore all arms uniformly
while true do
    if $t \in \mathcal{T}_2$ then
        if $|\ \widehat{\mu}_1(t) - \widehat{\mu}_2(t)\ | \geq \sqrt{16/t \log^+(T_1/t)}$ then
            break
        end if
    end if
    Pull arms $A_{t+1} = 1$ and $A_{t+2} = 2$, $t \leftarrow t + 2$
end while
Stage II: Commit to the arm with the largest average reward          (same as in Algorithm 2)
Stage III: Explore the unchosen arm in Stage II
$\mu' \leftarrow \widehat{\mu}_{1'}(t)$, $2' \leftarrow \{1,2\} \setminus 1'$, $t_2 \leftarrow 0$, $\theta_{2's}$ is the recalculated average reward of arm $2'$ after its
    s-th pull in Stage III and $\theta_{2's} \leftarrow 0$, for $s = 0$
while $t_2 \leq \log^2 T$ do
    if $t_2 \in \mathcal{T}_2'$ then
        if $|\mu' - \theta_{2',t_2}| < \sqrt{2/t_2 \log\big(T/t_2\big(\log^2(T/t_2) + 1\big)\big)}$ then
            break
        end if
    end if
    Pull arm $A_{t+1} = 2'$, $t \leftarrow t + 1$, $t_2 \leftarrow t_2 + 1$
end while
Stage IV: Commit to the arm with the largest average reward          (same as in Algorithm 2)
```

Theorem 3.3 *In the batched bandit problem, the expected number of rounds used in Algorithm 5 is $O(1)$. Moreover, the regret of Algorithm 5 is asymptotically optimal.*

The proof of Theorem 3.3 can be found in Section D.2. Here, we only focus on deriving the asymptotic optimality along with a constant round complexity in the batched bandits setting. For minimax and instance dependent regret bounds, Perchet et al. (2016) proved that any algorithm achieving the minimax optimality or instance dependent optimality will cost at least $\Omega(\log\log T)$ or $\Omega(\log T/\log\log T)$ rounds respectively. How to extending our minimax/instance-dependent and asymptotic optimal Algorithm 3 to the batched bandit setting is an interesting open question.

4. Experiment

In this section, we experimentally compare our proposed algorithms with existing algorithms including BAI-ETC (Garivier et al., 2016), SPRT-ETC (Garivier et al., 2016), and UCB (Garivier

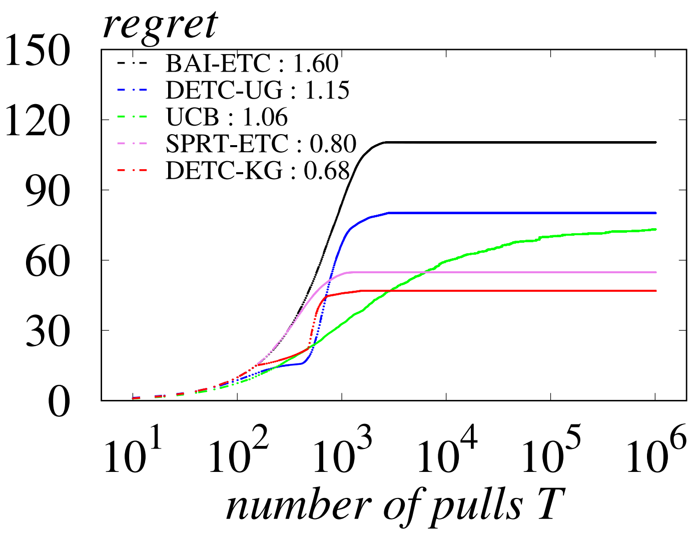

Figure 1: Regret comparison among ETC-type algorithms as well as UCB. Results are averaged over 10000 repetitions.

and Cappé, 2011). Our Algorithm 1 is denoted by DETC-UG and Algorithm 2 is denoted by DETC-KG. We test all the algorithms on a two-armed bandit with Gaussian rewards, where the mean reward follows distribution $\mathcal{N}(\mu_i, 1)$, for arm $i = 1, 2$. We set the gap between the two arms as $\Delta = 0.2$. For DETC-UG, the tuning parameter is T_1 and for DETC-KG, the tuning parameter is ϵ_T.

Among all compared algorithms, BAI-ETC, DETC-UG, and UCB are designed for the unknown gap setting while SPRT-ETC and DETC-KG are designed for the known gap setting (which use the gap information). All experiments are repeated 10000 times. Figure 1 shows the results on regret. In the legend, the regret rate $\Delta\widehat{R}(T)/\log T$ is appended after the algorithm names. As shown in Figure 1, the regret behavior reflects the theoretical results. In particular, DETC-UG achieves comparable regret with UCB and DETC-KG achieves the smallest regret. In addition, for both ETC and DETC strategies, the regret increases much slower after t exceeds certain threshold. The reason is that for both ETC and DETC strategies, the regret for the last exploitation stage is in the order of $O(1/\Delta) = O(1)$, which is a constant independent of T. This means most of the regret $O(\log T/\Delta)$ are due to the first three stages.

5. Conclusion

In this paper, we revisit the explore-then-commit (ETC) type of algorithms for multi-armed bandit problems, which separate the exploration and exploitation stages. We break the barrier that ETC type algorithms cannot achieve the asymptotically optimal regret bound (Garivier et al., 2016), which is usually attained by fully sequential strategies such as UCB. We propose a double explore-then-commit (DETC) strategy and prove that DETC is asymptotically optimal for subgaussian rewards, which is the first ETC type algorithm that matches the theoretical performance of UCB based algorithms. We also show a variant of DETC for two-armed bandit problems, which can achieve the asymptotic optimality and the minimax/instance-dependent regret bound simultaneously, while still keeping the merit of being non-fully-sequential.

To demonstrate the advantage of DETC over fully sequential strategies, we apply DETC to the batched bandit problem which has various real world applications and prove that DETC enjoys a constant round complexity while maintaining the asymptotic optimality at the same time. As a comparison, the round complexity of fully sequential strategies such as UCB usually scales with the horizon length T of the algorithm. This implies that the proposed DETC algorithm not only enjoys optimal regret bounds under various metrics, but is also practical and easily implementable in applications where the decision maker is expected to not switch its policy frequently and where the learning process is in a batch and parallel fashion.

Acknowledgments

X. Xiao is supported by the Ministry of Education, Singapore, under Tier-2 Grant R-252-000-A70-112. T. Jin is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-PhD/2021-01-004[T]). P. Xu and Q. Gu are partially supported by the National Science Foundation IIS-1904183. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Arpit Agarwal, Shivani Agarwal, Sepehr Assadi, and Sanjeev Khanna. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Conference on Learning Theory*, pages 39–75, 2017.

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

Miklos Ajtai, János Komlos, William L Steiger, and Endre Szemerédi. Deterministic selection in o (loglog n) parallel time. In *Proceedings of the eighteenth annual ACM symposium on Theory of computing*, pages 188–195, 1986.

Noga Alon and Yossi Azar. Sorting, approximate sorting, and searching in rounds. *SIAM Journal on Discrete Mathematics*, 1(3):269–280, 1988.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference On Learning Theory*, 2009.

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Dimitris Bertsimas and Adam J Mersereau. A learning approach for interactive marketing to a customer segment. *Operations Research*, 55(6):1120–1135, 2007.

Béla Bollobás and Andrew Thomason. Parallel sorting. *Discrete Applied Mathematics*, 6(1):1–11, 1983.

Mark Braverman, Jieming Mao, and S Matthew Weinberg. Parallel algorithms for select and partition with noisy comparisons. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 851–862, 2016.

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

Stephen E Chick and Noah Gans. Economic analysis of simulation selection problems. *Management Science*, 55(3):421–437, 2009.

Rémy Degenne and Vianney Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 1587–1595, 2016.

John Duchi, Feng Ruan, and Chulhee Yun. Minimax bounds on stochastic batched convex optimization. In *Conference On Learning Theory*, pages 3065–3162, 2018.

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

Hossein Esfandiari, Amin Karbasi, Abbas Mehrabian, and Vahab Mirrokni. Batched multi-armed bandits with optimal regret. *arXiv preprint arXiv:1910.04959*, 2019.

Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. In *Advances in Neural Information Processing Systems*, pages 501–511, 2019.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376, 2011.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027, 2016.

Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, pages 784–792, 2016.

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

Tianyuan Jin, Shi Jieming, Xiaokui Xiao, Enhong Chen, et al. Efficient pure exploration in adaptive round model. In *Advances in Neural Information Processing Systems*, pages 6605–6614, 2019.

Tianyuan Jin, Jing Tang, Pan Xu, Keke Huang, Xiaokui Xiao, and Quanqaun Gu. Almost optimal anytime algorithm for batched multi-armed bandits. In *International Conference on Machine Learning*, 2021.

Michael N Katehakis and Herbert Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92(19):8584, 1995.

Emilie Kaufmann et al. On bayesian index policies for sequential resource allocation. *The Annals of Statistics*, 46(2):842–865, 2018.

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

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

Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *The Journal of Machine Learning Research*, 19(1):765–796, 2018.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pages 223–237, 2017.

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

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

Chao Tao, Qin Zhang, and Yuan Zhou. Collaborative learning with limited interaction: Tight bounds for distributed exploration in multi-armed bandits. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 126–146. IEEE, 2019.

Leslie G Valiant. Parallelism in comparison problems. *SIAM Journal on Computing*, 4(3):348–355, 1975.

Jean Ville. *Étude critique de la notion de collectif*. 1939. URL http://eudml.org/doc/192893.

Contents

Appendix A. Related Work on Multi-Armed Bandits and Batched Bandits

For regret minimization in stochastic bandit problems, Lai and Robbins (1985) proved the first asymptotically lower bound that any strategy must have at least $C(\mu)\log(T)(1-o(1))$ regret when the horizon T approaches infinity, where $C(\mu)$ is a constant. Later, strategies such as UCB (Lai and Robbins, 1985; Auer et al., 2002a; Garivier and Cappé, 2011), Thompson Sampling (Korda et al., 2013; Agrawal and Goyal, 2017) and Bayes UCB (Kaufmann et al., 2018) are all shown to be asymptotically optimal in the unknown gap setting. For the known gap setting, Garivier et al. (2016) developed the Δ-UCB algorithm that matches the lower bound. To our knowledge, all previous asymptotically optimal algorithms are fully sequential. Despite the asymptotic optimality, for a fixed time horizon T, the problem-independent lower bound (Auer et al., 2002b) states that any strategy has at least a regret in the order of $\Omega(\sqrt{KT})$, which is called the *minimax optimal* regret. MOSS (Audibert and Bubeck, 2009) is the first method proved to be minimax optimal. Subsequently, two UCB-based methods, AdaUCB (Lattimore, 2018) and KL-UCB^{++} (Ménard and Garivier, 2017), are also shown to achieve minimax optimality.

There is less work yet focusing on the batched bandit setting with limited rounds. UCB2 (Auer et al., 2002a), which needs implicitly $O(\log T)$ rounds of queries, is a variant of UCB that takes $O(T)$ rounds of queries. Cesa-Bianchi et al. (2013) studied the batched bandit problem under the notion of switching cost and showed that $\log\log T$ rounds are sufficient to achieve the minimax optimal regret Audibert and Bubeck (2009). Perchet et al. (2016) studied the two-armed batched bandit problem with limited rounds. They developed polices that is minimax optimal and proved that their round cost is near optimal. Gao et al. (2019) used similar polices for K-armed batched bandits and proved that their batch complexity and regret are both near optimal, which is recently further improved by Esfandiari et al. (2019). Besides, Gao et al. (2019); Esfandiari et al. (2019); Perchet et al. (2016) also provide the instance dependent regret bound under the limited rounds setting. In the asymptotic sense, the regret bound is $O(K\log T)$. However, the hidden constant in $O(K\log T)$ makes it suboptimal in terms of the asymptotic regret and the round cost of these works is $\Theta(\log T)$. In addition, the batched bandit problem is also studied in the linear bandit setting (Esfandiari et al., 2019; Han et al., 2020; Ruan et al., 2020), best arm identification Agarwal et al. (2017); Jin et al. (2019) and in theoretical computer science under the name of *parallel algorithms* (Valiant, 1975; Tao et al., 2019; Alon and Azar, 1988; Feige et al., 1994; Bollobás and Thomason, 1983; Ajtai et al., 1986; Braverman et al., 2016; Duchi et al., 2018), to mention a few.

Appendix B. An Anytime Algorithm with Asymptotic Optimality

In previous sections, the stopping rules of DETC depends on the horizon length T. However, this may not be the case in some practical cases, where we prefer to stop the algorithm at an arbitrary time without deciding it at the beginning. This is referred to as the anytime setting in the bandit literature (Degenne and Perchet, 2016; Lattimore and Szepesvári, 2020). In this section, we provide an extension of our DETC algorithm for two-armed bandits to the anytime setting. Our algorithm guesses T in epochs. For the r-th epoch, we guess $T = 2^{r+1}$. At the r-th epoch of the algorithm, the algorithm proceeds as follows: we find the arm $1'$ which is the arm that played most often in the first $r-1$ epochs; then we pull arm $2'$ till the stopping rules

$$|\widehat{\mu}_{1'}(t) - \widehat{\mu}_{2'}(t)| < \sqrt{\frac{2}{T_{2'}(t)}\log\left(\frac{r\cdot 2^r}{T_{2'}(t)}\left(\log^2\left(\frac{r\cdot 2^r}{T_{2'}(t)}\right)+1\right)\right)} \qquad \text{and} \qquad t \le 2^{r+1}$$

Algorithm 6 Anytime Asymptotically Optimal ETC in the Unknown Gap Setting

```
**Initialization:** Pull arms $A_1 = 1$, $A_2 = 2$, $t \leftarrow 2$;
**for** $r = 1, 2, \cdots$ **do**
    $1' \leftarrow \arg\max_{i \in \{1,2\}} T_i(t)$, $2' \leftarrow \{1, 2\} \setminus 1'$
    **while** $|\widehat{\mu}_{1'}(t) - \widehat{\mu}_{2'}(t)| < \sqrt{\frac{2}{T_{2'}(t)} \log\big(\frac{r \cdot 2^r}{T_{2'}(t)}\big(\log^2(\frac{r \cdot 2^r}{T_{2'}(t)}) + 1\big)\big)}$ and $t \leq 2^{r+1}$ **do**
        $A_{t+1} = 2'$, $t \leftarrow t + 1$
    **end while**
    $a(r) \leftarrow 1' \, \mathbb{1}\{\widehat{\mu}_{1'}(t) \geq \widehat{\mu}_{2'}(t)\} + 2' \, \mathbb{1}\{\widehat{\mu}_{1'}(t) < \widehat{\mu}_{2'}(t)\}$
    **while** $t \leq 2^{r+1}$ **do**
        Pull arm $a(r)$, $t \leftarrow t + 1$
    **end while**
**end for**
```

is satisfied. The aim here is to ensures that the difference of the average reward between the arm $2'$ and $1'$ is sufficient large such that the regret of playing the winner of $1'$ and $2'$ 2^r times is bounded. As we will prove: we keep the optimal regret by sightly improve the term in the stopping condition from 2^r to $r \cdot 2^r$. After this step, we commit to the arm with the largest average reward. Compared with Algorithm 2, in each epoch, anytime ETC seems only to perform the third stage and fourth stage of Algorithm 2. The reason here is that: the first two stages of Algorithm 2 aims to pull one arm $\log^2 T$ times while keeping the optimal regret. In anytime ETC algorithm, when the algorithm runs $\log^2 T$ steps, $1'$ is the arm that pulled most often, thus $1'$ is pulled $O(\log^2 T)$ times. Besides, as we will prove later that the regret of first $\log^2 T$ steps is bounded by $O(\sqrt{\log T})$.

The following theorem shows that Algorithm 6 is still asymptotically optimal for an unknown horizon T.

Theorem B.1 *The total expected regret for the anytime version of DETC (Algorithm 6) satisfies* $\lim_{T \to \infty} R_\mu(T)/\log T = 2/\Delta$.

The proof of Theorem B.1 could be found in Section C.4. The result shows that even for the anytime setting (unknown horizon length), the ETC strategy can also be asymptotically optimal as the UCB (Katehakis and Robbins, 1995) and Thompson Sampling (Korda et al., 2013) do. An advantage of the anytime ETC algorithm is that Algorithm 6 only needs $O(\log T)$ epochs, where in each epoch it can separate exploration and exploitation stages, while for anytime UCB or Thompson Sampling algorithms often need $O(T)$ mixed exploration and exploitation stages.

Appendix C. Double Explore-then-Commit for K-Armed Bandits

In this section, we extend our DETC framework to K-armed bandit problems, where $K > 2$. Due to the similarity in both structures and analyses between Algorithm 1 for the known gap setting and Algorithm 2 for the unknown gap setting, we only present the K-armed bandit algorithm for the unknown gap setting, which is usually more general in practice and challenging in analysis.

We present the double explore-then-commit algorithm for K-armed bandits in Algorithm 7. Similar to Algorithm 2 for two-armed bandits, the algorithm proceeds as follows: (1) in *Stage I*, we uniformly explore over all the K arms; (2) in *Stage II*, we pull the arm with the largest average

Algorithm 7 Double Explore-then-Commit for K-Armed Bandits (DETC-K)

```
input $T$, $K$. Initialization: $t \leftarrow 0$
  Stage I: Explore all arms uniformly
  while $t \le K\sqrt{\log T}$ do
    Pull every arm once, $t \leftarrow t + K$
  end while
  Stage II: Commit to the arm with the largest average reward
  $1' \leftarrow \arg\max_k \widehat{\mu}_k(t)$, $s \leftarrow 0, p_0 \leftarrow 0$
  while $s \le \log^2 T$ do
    Pull arm $A_{t+1} = 1'$ and observe reward $r_{t+1}$
    $p_{s+1} = (s \cdot p_s + r_{t+1})/(s+1)$, $s \leftarrow s+1, t \leftarrow t+1$
  end while
  Stage III: Explore the unchosen arm in Stage II
  $\mu' \leftarrow p_s$, Denote $\{2', \cdots, K'\} = \{1, 2, \cdots, K\} \setminus \{1'\}$
  for $i = 2, 3, \cdots, K$ do
    $t_i \leftarrow 1, \theta_{i',0} = 0$
    while $|\mu' - \theta_{i',t_i}| < \sqrt{2/t_i \log\big(T/t_i\big(\log^2(T/t_i) + 1\big)\big)}$ and $t_i \le \log^2 T$ do
      Pull arm $i'$ and observe reward $r_{t+1}$
      $\theta_{i',t_i+1} = (t_i \cdot \theta_{i',t_i} + r_{t+1})/(t_i + 1)$, $t \leftarrow t+1$, $t_i \leftarrow t_i + 1$
    end while
    if $t_i > \log^2 T$ then
      $\mathcal{F}_{\text{fail}} \leftarrow 1$ and break
    end if
  end for
  Stage IV: Commit to the arm with the largest average reward
  $j' := \max_{i'} \theta_{i't_i}$
  if $\widehat{\mu}_{1'} \ge \theta_{j't_j}$ and $\mathcal{F}_{\text{fail}} = 0$ then
    Let $a \leftarrow 1'$
    while $t < T$ do
      Pull arm $a$, $t \leftarrow t+1$
    end while
  else
    Pull every arm $\log^2 T$ times and let $a$ be the arm with the largest average reward for this pull
    Pull arm $a$ till $T$ time steps.
  end if
```

reward; (3) in *Stage III*, we aim to ensure that the difference between the chosen arm $1'$ in *Stage II* and unchosen arms is sufficient by pulling all the unchosen arm i' ($i \ge 2$) repeatedly until the average reward of arm i' collected in this stage can be clearly distinguished from the average reward of arm $1'$. We set a check flag $\mathcal{F}_{\text{fail}}$ initialized as 0, which will be set to 1 if any unchosen arm i' is pulled for $\log^2 T$ times; (4) in *Stage IV*, if $\mathcal{F}_{\text{fail}} = 0$ and $\widehat{\mu}_{1'}$ is larger than the recalculated average reward for any other arm, then we pull $1'$ till the end. Otherwise, $1'$ may not be the best arm. Then we pull all arms $\log^2 T$ times, and pull the arm with the largest recalculated average reward till the end.

Now we present the regret bound of Algorithm 7.

Theorem C.1 *The regret of Algorithm 7 with 1-subgaussian rewards satisfies*

$$\lim_{T \to \infty} R_\mu(T)/\log(T) = \textstyle\sum_{i:\Delta_i > 0} 2/\Delta_i. \tag{C.1}$$

The proof of Theorem C.1 can be found in Section C.5. In the second case of *Stage IV* of Algorithm 7, we actually believe that we have failed to choose the best arm via previous stages and need to explore again for a fixed number of pulls ($\log^2 T$) for all arms and commit to the best arm based on the pulling results. Note that this can be seen as the naive ETC strategy with fixed design (Garivier et al., 2016), which has an asymptotic regret rate $4/\Delta$. Fortunately, Theorem C.1 indicates our DETC algorithm can still achieve the asymptotically optimal regret for K-armed bandits (Lai and Robbins, 1985). This means that the probability of failing in the first three stages of Algorithm 7 is rather small and thus the extra ETC step does not affect the asymptotic regret of our DETC algorithm. Lastly, it would be an interesting problem to extend the idea of Algorithm 3 in two-armed bandits to K-armed bandits, where simultaneously achieving the instance-dependent and asymptotically optimal regret is still an open problem (Agrawal and Goyal, 2017; Lattimore, 2018).

C.1. Proof of the Regret Bound of Algorithm 1

Now we are going to prove Theorem 2.1. We first present a technical lemma that characterizes the concentration properties of subgaussian random variables.

Lemma C.2 (Corollary 5.5 in Lattimore and Szepesvári (2020)) *Assume that $X_1, \ldots, X_n$ are independent, σ-subguassian random variables centered around μ. Then for any $\epsilon > 0$*

$$\mathbb{P}(\widehat{\mu} \geq \mu + \epsilon) \leq \exp\bigg(-\frac{n\epsilon^2}{2\sigma^2}\bigg) \quad \textit{and} \quad \mathbb{P}(\widehat{\mu} \leq \mu - \epsilon) \leq \exp\bigg(-\frac{n\epsilon^2}{2\sigma^2}\bigg), \tag{C.2}$$

where $\widehat{\mu} = 1/n\sum_{t=1}^n X_t$.

Proof [Proof of Theorem 2.1] Let τ_2 be the total number of times arm $2'$ is pulled in *Stage III* of Algorithm 1. We know that τ_2 is a random variable. Recall that $\mu_1 > \mu_2$ and $\Delta = \mu_1 - \mu_2$. Recall τ_1 is number of times arm 1 is pulled in *Stage I*. Let $N_2(T)$ denote the total number of times Algorithm 1 pulls arm 2, which is calculated as

$$\begin{aligned} N_2(T) = \tau_1 + (T_1 - \tau_1)\,\mathbb{1}\{\widehat{\mu}_1(\tau_1) < \widehat{\mu}_2(\tau_1)\} + \tau_2\,\mathbb{1}\{\widehat{\mu}_1(\tau_1) \geq \widehat{\mu}_2(\tau_1)\} \\ + (T - T_1 - \tau_1 - \tau_2)\,\mathbb{1}\{a = 2\}. \end{aligned} \tag{C.3}$$

Then, the regret of Algorithm 1 $R_\mu(T) = \mathbb{E}[\Delta N_2(T)]$ can be decomposed as follows

$$\begin{aligned} R_\mu(T) &\leq \mathbb{E}\big[\Delta\tau_1 + \Delta(T_1 - \tau_1)\,\mathbb{1}\{\widehat{\mu}_1(\tau_1) < \widehat{\mu}_2(\tau_1)\} + \Delta\tau_2\,\mathbb{1}\{\widehat{\mu}_1(\tau_1) \geq \widehat{\mu}_2(\tau_1)\,\mathbb{1}\} + \Delta T\,\mathbb{1}\{a = 2\}\big] \\ &\leq \mathbb{E}\big[\Delta\tau_1 + \Delta T_1\mathbb{P}(\widehat{\mu}_1(\tau_1) < \widehat{\mu}_2(\tau_1)) + \Delta\tau_2\mathbb{P}(\widehat{\mu}_1(\tau_1) \geq \widehat{\mu}_2(\tau_1)) + \Delta T\mathbb{P}(a = 2)\big] \\ &\leq \Delta\tau_1 + \underbrace{\Delta T_1\mathbb{P}(\tau_1 < T_1, 1' = 2)}_{I_1} + \underbrace{\Delta\mathbb{E}[\tau_2]}_{I_2} + \underbrace{\Delta T\mathbb{P}(\tau_2 < T, a = 2)}_{I_3}. \end{aligned} \tag{C.4}$$

In what follows, we will bound these terms separately.

Bounding term I_1: Let X_i and Y_i be the rewards from pulling arm 1 and arm 2 for the i-th time respectively. Thus $X_i - \mu_1$ and $Y_i - \mu_2$ are 1-subgaussian random variables. Let $S_0 = 0$ and $S_n = (X_1 - Y_1) + \cdots + (X_n - Y_n)$ for every $n \geq 1$. Then $X_i - Y_i - \Delta$ is a $\sqrt{2}$-subgaussian random variable. Applying Lemma C.2 with any $\epsilon > 0$, we get

$$\mathbb{P}(S_{\tau_1}/\tau_1 \leq \Delta - \epsilon) \leq \exp(-\tau_1\epsilon^2/4) \leq \exp(-\epsilon^2\log(T_1\Delta^2)/\Delta^2), \tag{C.5}$$

where in the last inequality we plugged in the fact that $\tau_1 \geq 4\log(T_1\Delta^2)/\Delta^2$. By setting $\epsilon = \Delta$ in the above inequality, we further obtain $\mathbb{P}(\tau_1 < T_1, 1' = 2) = \mathbb{P}(S_{\tau_1}/\tau_1 \leq 0) \leq 1/(T_1\Delta^2)$. Hence

$$I_1 = T_1\Delta\mathbb{P}(\tau_1 < T_1, 1' = 2) \leq 1/\Delta. \tag{C.6}$$

Bounding term I_2: Recall that $T_1 \geq 2\log(T\Delta^2)/(\epsilon_T^2\Delta^2)$. Define event $E = \{\mu' \in (\mu_{1'} - \epsilon_T\Delta, \mu_{1'} + \epsilon_T\Delta)\}$, and let E^c be the complement of E. By Lemma C.2 and the union bound, $\mathbb{P}(E) \geq 1 - 2/(T\Delta^2)$. Therefore,

$$\begin{aligned} I_2 &= \Delta\mathbb{E}[\tau_2 \mathbb{1}(E)] + \Delta\mathbb{E}[\tau_2 \mathbb{1}(E^c)] \\ &= \Delta\mathbb{E}[\tau_2 \mathbb{1}(E)] + \Delta\mathbb{E}[\tau_2 \mid E^c]\cdot\mathbb{P}(E^c) \\ &\leq \Delta\mathbb{E}[\tau_2 \mathbb{1}(E)] + \Delta T\cdot\frac{2}{T\Delta^2} \\ &= \Delta\mathbb{E}[\tau_2 \mathbb{1}(E, 1' = 1)] + \Delta\mathbb{E}[\tau_2 \mathbb{1}(E, 1' = 2)] + 2/\Delta. \end{aligned} \tag{C.7}$$

We first focus on term $\Delta\mathbb{E}[\tau_2 \mathbb{1}(E, 1' = 1)]$. Observe that when E holds and $1' = 1$ (i.e., the chosen arm $1'$ is the best arm), arm $2' = 2$ is pulled in *Stage III* of Algorithm 1. For ease of presentation, we define the following notations:

$$Z_0 = 0, \quad Z_i = \mu' - Y_{i+\tau_1}, \quad S'_0 = 0, \quad S'_n = Z_1 + \cdots + Z_n, \tag{C.8}$$

where $Y_{i+\tau_1}$ is the reward from pulling arm 2 for the *i-th* time in Stage *III*. For any $x > 0$, we define

$$n_x = (\log(T\Delta^2) + x)/(2(1 - \epsilon_T)^2\Delta^2).$$

We also define a check point parameter $x_0 = 2\sqrt{\log(T\Delta^2)}$.

Let E_1 denote the event $\{E, 1' = 1\}$. Note that in *Stage III* of Algorithm 1, conditioned on E_1, we have

$$2(1 - \epsilon_T)\Delta|S'_{t_2}| = 2(1 - \epsilon_T)t_2\Delta|\mu' - \theta_{2',t_2}| < \log(T\Delta^2),$$

for $t_2 \leq \tau_2 - 1$. Therefore, conditioned on E_1,

$$\begin{aligned} \left\{\tau_2 - 1 \geq \left\lceil\frac{\log(T\Delta^2) + x}{2(1 - \epsilon_T)^2\Delta^2}\right\rceil\right\} &= \{\tau_2 - 1 \geq \lceil n_x\rceil\} \\ &\subseteq \left\{S'_{\lceil n_x\rceil} \leq \frac{\log(T\Delta^2)}{2(1 - \epsilon_T)\Delta}\right\}. \end{aligned} \tag{C.9}$$

Let $\Delta' = \mu' - \mathbb{E}[Y_{i+\tau_1}]$. Then, $Z_i - \Delta'$ is 1-subgaussian. We have that conditioned on E_1,

$$\Delta' = \mu' - \mathbb{E}[Y_{1+\tau_1}] = \mu' - \mu_2 \geq \mu_1 - \epsilon_T\Delta - \mu_2 = (1 - \epsilon_T)\Delta. \tag{C.10}$$

By Lemma C.2, for any $\epsilon > 0$, we have

$$\mathbb{P}\left(\frac{S'_{\lceil n_x\rceil}}{\lceil n_x\rceil} \leq \Delta' - \epsilon \,\middle|\, E_1\right) \leq \exp\left(-\lceil n_x\rceil\epsilon^2/2\right). \tag{C.11}$$

Let $\epsilon = \frac{(1-\epsilon_T)\Delta x}{\log(T\Delta^2)+x}$. Conditioned on E_1,

$$\lceil n_x \rceil (\Delta' - \epsilon) \geq \lceil n_x \rceil ((1-\epsilon_T)\Delta - \epsilon) \geq \frac{\log(T\Delta^2)}{2(1-\epsilon_T)\Delta}.$$

Combining this with (C.11) yields

$$\begin{aligned} \mathbb{P}\left(S'_{\lceil n_x \rceil} \leq \frac{\log(T\Delta^2)}{2(1-\epsilon_T)\Delta} \;\middle|\; E_1 \right) &\leq \mathbb{P}\left(S'_{\lceil n_x \rceil} \leq \lceil n_x \rceil (\Delta' - \epsilon) \;\middle|\; E_1 \right) \\ &\leq \exp\left(- \frac{x^2}{4(\log(T\Delta^2)+x)} \right). \end{aligned} \tag{C.12}$$

This, when combined with (C.9), implies

$$\mathbb{P}\left(\tau_2 - 1 \geq \left\lceil \frac{\log(T\Delta^2)+x}{2(1-\epsilon_T)^2\Delta^2} \right\rceil \;\middle|\; E_1 \right) \leq \exp\left(- \frac{x^2}{4(\log(T\Delta^2)+x)} \right).$$

Recall that $x_0 = 2\sqrt{\log(T\Delta^2)}$. For any $x \geq x_0$, we have $x\sqrt{\log(T\Delta^2)}/2 \geq \log(T\Delta^2)$. Thus,

$$\begin{aligned} \int_{n_{x_0}}^{\infty} \mathbb{P}(\tau_2 - 2 \geq v \mid E_1) \mathrm{d}v &= \int_{x_0}^{\infty} \mathbb{P}\left(\tau_2 - 2 \geq \frac{\log(T\Delta^2)+x}{2(1-\epsilon_T)^2\Delta^2} \;\middle|\; E_1 \right) \frac{\mathrm{d}x}{2(1-\epsilon_T)^2\Delta^2} \\ &\leq \int_{x_0}^{\infty} \mathbb{P}\left(\tau_2 - 1 \geq \left\lceil \frac{\log(T\Delta^2)+x}{2(1-\epsilon_T)^2\Delta^2} \right\rceil \;\middle|\; E_1 \right) \frac{\mathrm{d}x}{2(1-\epsilon_T)^2\Delta^2} \\ &\leq \frac{1}{2(1-\epsilon_T)^2\Delta^2} \int_{x_0}^{\infty} \exp\left(- \frac{x^2}{4(\log(T\Delta^2)+x)} \right) \mathrm{d}x \\ &\leq \frac{1}{2(1-\epsilon_T)^2\Delta^2} \int_{x_0}^{\infty} \exp\left(- \frac{x}{2\sqrt{\log(T\Delta^2)}+4} \right) \mathrm{d}x \\ &\leq \frac{1}{2(1-\epsilon_T)^2\Delta^2} \int_{0}^{\infty} \exp\left(- \frac{x}{2\sqrt{\log(T\Delta^2)}+4} \right) \mathrm{d}x \\ &= \frac{\sqrt{\log(T\Delta^2)}+2}{(1-\epsilon_T)^2\Delta^2}. \end{aligned} \tag{C.13}$$

Then, the expectation of $\Delta\tau_2$ conditioned on E_1 is

$$\begin{aligned} \Delta\mathbb{E}[\tau_2 \mid E_1] &= \Delta \int_0^{\infty} \mathbb{P}(\tau_2 > v \mid E_1) \mathrm{d}v \\ &= \Delta \int_0^{n_{x_0}+2} \mathbb{P}(\tau_2 > v \mid E_1) \mathrm{d}v + \Delta \int_{n_{x_0}}^{\infty} \mathbb{P}(\tau_2 - 2 \geq v \mid E_1) \mathrm{d}v \\ &\leq 2\Delta + \frac{\log(T\Delta^2)}{2(1-\epsilon_T)^2\Delta} + \frac{2\sqrt{\log(T\Delta^2)}+2}{(1-\epsilon_T)^2\Delta}. \end{aligned} \tag{C.14}$$

Hence, we have

$$\begin{aligned} \Delta\mathbb{E}[\tau_2 \mathbb{1}(E, 1' = 1)] &= \Delta\mathbb{E}[\tau_2 \mid E_1] \cdot \mathbb{P}(E_1) \\ &\leq \mathbb{P}(E_1) \cdot \left(2\Delta + \frac{\log(T\Delta^2)}{2(1-\epsilon_T)^2\Delta} + \frac{2\sqrt{\log(T\Delta^2)}+2}{(1-\epsilon_T)^2\Delta} \right). \end{aligned} \tag{C.15}$$

Let E_2 denote the event $\{E, 1' = 2\}$. In a manner similar to the proof of (C.14), we can show that

$$\begin{aligned}\Delta\mathbb{E}[\tau_2 \mathbb{1}(E, 1' = 2)] &= \Delta\mathbb{E}[\tau_2 \mid E_2] \cdot \mathbb{P}(E_2) \\ &\leq \mathbb{P}(E_2) \cdot \bigg(2\Delta + \frac{\log(T\Delta^2)}{2(1-\epsilon_T)^2\Delta} + \frac{2\sqrt{\log(T\Delta^2)} + 2}{(1-\epsilon_T)^2\Delta}\bigg). \qquad \text{(C.16)}\end{aligned}$$

Therefore, we have

$$\begin{aligned}I_2 &\leq \Delta\mathbb{E}[\tau_2 \mathbb{1}(E, 1' = 1)] + \Delta\mathbb{E}[\tau_2 \mathbb{1}(E, 1' = 2)] + \frac{2}{\Delta} \\ &\leq 2\Delta + \frac{2}{\Delta} + \frac{\log(T\Delta^2)}{2(1-\epsilon_T)^2\Delta} + \frac{2\sqrt{\log(T\Delta^2)} + 2}{(1-\epsilon_T)^2\Delta}. \qquad \text{(C.17)}\end{aligned}$$

Bounding term I_3: For term I_3, similar to (C.7), we have

$$\begin{aligned}I_3 =& \Delta \cdot T\mathbb{P}[\tau_2 < T, a = 2 \mid E_1] \cdot \mathbb{P}[E_1] \\ &+ \Delta \cdot T\mathbb{P}[\tau_2 < T, a = 2 \mid E_2] \cdot \mathbb{P}[E_2] + \frac{2}{\Delta}. \qquad \text{(C.18)}\end{aligned}$$

We will first prove that $\mathbb{P}(\tau_2 < T, a = 2 \mid E_1) \leq 1/(T\Delta^2)$. Recall that $S'_n = \sum_{i=1}^n Z_i$ and $Z_i = \mu' - Y_{i+\tau_1}$. In addition, $Z_i - \Delta'$ is 1-subgaussian, and $\Delta' \geq (1-\epsilon_T)\Delta$ whenever E_1 occurs. Then,

$$\begin{aligned}\mathbb{E}[\exp(-2\Delta(1-\epsilon_T)Z_1) \mid E_1] &= \mathbb{E}[\exp(-2\Delta(1-\epsilon_T)Z_1 + 2\Delta\Delta'(1-\epsilon_T) - 2\Delta\Delta'(1-\epsilon_T)) \mid E_1] \\ &= \mathbb{E}[\exp(-2\Delta(1-\epsilon_T)(Z_1 - \Delta') - 2\Delta\Delta'(1-\epsilon_T)) \mid E_1] \\ &\leq \exp((-2(1-\epsilon_T)\Delta)^2/2 - 2(1-\epsilon_T)\Delta\Delta')) \\ &\leq \exp(2(1-\epsilon_T)\Delta((1-\epsilon_T)\Delta - \Delta')) \\ &\leq 1, \qquad \text{(C.19)}\end{aligned}$$

where the first inequality follows from the definition of subgaussian random variables. We consider the sigma-algebra $F_n = \sigma(E_1, Y_{\tau_1+i}, i = 1, ..., n)$ for $n \geq 1$. Define $F_0 = E_1$ and $M_0 = 1$. Then, the sequence $\{M_n\}_{n=0,1,...}$ with $M_n = \exp(-2\Delta(1-\epsilon_T)S'_n)$ is a super-martingale with respect to $\{F_n\}_{n=0,1,...}$. Let $\tau' = T \wedge \inf\{n > 1 : S'_n \leq -\log(T\Delta^2)/(2\Delta(1-\epsilon_T))\}$ be a stopping time. Observe that conditioned on E_1,

$$\begin{aligned}\{\tau_2 < T, a = 2\} &\subseteq \bigg\{\exists 1 < n < T : S'_n \leq -\frac{\log(T\Delta^2)}{2\Delta(1-\epsilon_T)}\bigg\} \\ &= \{\tau' < T\}. \qquad \text{(C.20)}\end{aligned}$$

Applying Doob's optional stopping theorem (Durrett, 2019) yields $\mathbb{E}[M_{\tau'}] \leq \mathbb{E}[M_0] = 1$. In addition, when $\tau_2 < T$, we have

$$\begin{aligned}M_{\tau'} &= \exp(-2\Delta(1-\epsilon_T)S'_{\tau'}) \\ &\geq \exp(\log(T\Delta^2)) = T\Delta^2. \qquad \text{(C.21)}\end{aligned}$$

In other words, $\{\tau_2 < T\} \subseteq \{M_{\tau'} \geq T\Delta^2\}$. This leads to

$$\begin{aligned} \mathbb{P}(\tau_2 < T, a = 2 \mid E_1) &\leq \mathbb{P}(\tau' < T \mid E_1) \\ &\leq \mathbb{P}(M_{\tau'} \geq T\Delta^2 \mid E_1) \\ &\leq \mathbb{E}[M_{\tau'}]/(T\Delta^2) \leq 1/(T\Delta^2). \end{aligned} \tag{C.22}$$

where the third inequality follows form Markov's inequality. Similarly, $\mathbb{P}(\tau_2 < T, a = 2 \mid E_2) \leq 1/(T\Delta^2)$ also holds. Thus, term I_3 can be upper bounded by $3/\Delta$.

Completing the proof: Substituting (C.6), (C.17) and $I_3 \leq 3/\Delta$ into (C.4) yields a total regret as follows

$$R_\mu(T) \leq 2\Delta + \frac{8}{\Delta} + \frac{4\log(T_1\Delta^2)}{\Delta} + \frac{\log(T\Delta^2) + 2\sqrt{\log(T\Delta^2)}}{2(1-\epsilon_T)^2\Delta} + \frac{\sqrt{\log(T\Delta^2)} + 2}{(1-\epsilon_T)^2\Delta}.$$

Recall the choice of ϵ_T in Theorem 2.1. By our choice that $T_1 = \lceil 2\log(T\Delta^2)/(\epsilon_T^2\Delta^2))\rceil$, we have

$$T_1 \leq 1 + \max\{2\log^2 T, 8\log(T\Delta^2)/\Delta^2\}, \tag{C.23}$$

which immediately implies, $\lim_{T\to\infty} 4\text{log}(T_1\Delta^2)/(\Delta\log T) = 0$. Also note that $\lim_{T\to\infty}\epsilon_T = 0$. Thus, we have $\lim_{T\to\infty} R_\mu(T)/\log T = 1/(2\Delta)$. By (C.23), we known that $T_1\Delta^2 = O(\log(T\Delta^2))$, which results in the worse case regret bound as

$$R_\mu(T) = O\bigg(\Delta + \frac{1}{\Delta} + \frac{\log(T\Delta^2)}{\Delta} + \frac{\log\log(T\Delta^2)}{\Delta}\bigg) = O(\Delta + \frac{\log(T\Delta^2)}{\Delta}) = O(\Delta + \sqrt{T}),$$

where the last equality is due to the fact that $T\Delta^2 > 1$ and $\log x \leq 2\sqrt{x}$ for $x > 1$.

■

C.2. Proof of the Regret Bound of Algorithm 2

Next, we provide the proof for Theorem 2.2. Note that the stopping time of *Stage I* and *Stage III* in Algorithm 2 is not fixed and instead depends on the random samples, and hence, the Hoeffding's inequality in Lemma C.2 is not directly applicable. To address this issue, we provide the following two Lemmas.

Lemma C.3 *Let N and M be extended real numbers in $\mathbb{R}^+$ and $\mathbb{R}^+ \cup \{+\infty\}$. Let γ be a real number in $\mathbb{R}^+$, and let $\widehat{\mu}_n = \sum_{s=1}^n X_s/n$ be the empirical mean of n random variables identically independently distributed according to 1-subgaussian distribution. Then*

$$\mathbb{P}(\exists N \leq n \leq M, \widehat{\mu}_n + \gamma \leq 0) \leq \exp\bigg(-\frac{N\gamma^2}{2}\bigg). \tag{C.24}$$

The following lemma characterizes the length of the uniform exploration in *Stage I* of Algorithm 2. Since each arm is pulled for the same number of times (e.g., s times), the length of *Stage I* is $2s$.

Lemma C.4 *Let $n \in \mathbb{N}^+$, $X_1, X_2, \cdots$, be i.i.d. 1-subgaussian random variables, and $Y_1, Y_2, \cdots$, be i.i.d. 1-subgaussian random variables. Assume without loss of generality that $\mathbb{E}[X_1] > \mathbb{E}[Y_1]$. Denote $\Delta = \mathbb{E}[X_i - Y_i]$, and $\widehat{\mu}_t = 1/\sum_{n=1}^{t}(X_n - Y_n)$. Then for any $x > 0$,*

$$\mathbb{P}\bigg(\exists s \geq 1 : \widehat{\mu}_s + \sqrt{\frac{8}{s}\log^+\bigg(\frac{N}{s}\bigg)} \leq 0\bigg) \leq \frac{15}{N\Delta^2}.$$

Moreover, we need following inequalities on the confidence bound of the average rewards. Similar results have also been proved in Ménard and Garivier (2017) for bounding the KL divergence between two exponential family distributions for different arms.

Lemma C.5 *Let $\delta > 0$ and $M_1, M_2, \ldots, M_n$ be 1-subgaussian random variables with zero means. Denote $\widehat{\mu}_n = \sum_{s=1}^{n} M_s/n$. Then the following statements hold:*

1. *for any $T_1 \leq T$,*

$$\sum_{n=1}^{T}\mathbb{P}\bigg(\widehat{\mu}_n + \sqrt{\frac{4}{n}\log^+\bigg(\frac{T_1}{n}\bigg)} \geq \delta\bigg) \leq 1 + \frac{4\log^+(T_1\delta^2)}{\delta^2} + \frac{3}{\delta^2} + \frac{\sqrt{8\pi\log^+(T_1\delta^2)}}{\delta^2}; \tag{C.25}$$

2. *if $T\delta^2 \geq e^2$, then*

$$\begin{aligned}&\sum_{n=1}^{T}\mathbb{P}\Bigg(\widehat{\mu}_n + \sqrt{\frac{2}{n}\log\bigg(\frac{T}{n}\bigg(\log^2\frac{T}{n} + 1\bigg)\bigg)} \geq \delta\Bigg)\\ &\leq 1 + \frac{2\log(T\delta^2(\log^2(T\delta^2)+1))}{\delta^2} + \frac{3}{\delta^2} + \frac{\sqrt{4\pi\log(T\delta^2(\log^2(T\delta^2)+1))}}{\delta^2};\end{aligned} \tag{C.26}$$

3. *if $T\delta^2 \geq 4e^3$, then*

$$\mathbb{P}\Bigg(\exists s \leq T : \widehat{\mu}_s + \sqrt{\frac{2}{s}\log\bigg(\frac{T}{s}\bigg(\log^2\frac{T}{s} + 1\bigg)\bigg)} + \delta \leq 0\Bigg) \leq \frac{4(16e^2+1)}{T\delta^2}. \tag{C.27}$$

Proof [Proof of Theorem 2.2] Let τ_1 be the number of times each arm is pulled in *Stage I* of Algorithm 2 and τ_2 be the total number of times arm $2'$ is pulled in *Stage III* of Algorithm 2. Similar to (C.4), the regret of Algorithm 2 can be decomposed as follows

$$R_\mu(T) \leq \underbrace{\Delta T_1\mathbb{P}(\tau_1 < T, 1' = 2)}_{I_1} + \underbrace{\Delta\mathbb{E}[\tau_1] + \Delta\mathbb{E}[\tau_2]}_{I_2} + \underbrace{\Delta T\mathbb{P}(\tau_2 < T, a = 2)}_{I_3}. \tag{C.28}$$

Since we focus on the asymptotic optimality, we define $\epsilon_T = \sqrt{2\log(T\Delta^2)/(T_1\Delta^2)}$ and assume $\epsilon_T \in (0, 1/2)$, $T\Delta^2 \geq 16e^3$.

Bounding term I_1: Let X_s and Y_s be the reward of arm 1 and 2 when they are pulled for the s-th time respectively, $s = 1, 2, \ldots$. Recall that $\widehat{\mu}_{k,s}$ is the average reward for arm k after its s-th pull. Applying Lemma C.4, we have

$$\mathbb{P}(\tau_1 < T, 1' = 2) \leq \mathbb{P}\Bigg(\exists s \in \mathbb{N} : 2s \leq T,\ \widehat{\mu}_{1,s} - \widehat{\mu}_{2,s} \leq -\sqrt{\frac{8\log^+(T_1/(2s))}{s}}\Bigg)$$

$$\le \frac{30}{T_1\Delta^2}. \tag{C.29}$$

where the last inequality comes from Lemma C.4. Therefore $I_1 \le 30/\Delta$.

Bounding term I_2: By the definition of τ_1 and the stopping rule of *Stage I* in Algorithm 2, we have

$$\begin{aligned}
\mathbb{E}[\tau_1] = \sum_{s=1}^{T} \mathbb{P}(\tau_1 \ge s) &\le \sum_{s=1}^{T/2} \mathbb{P}\bigg(\widehat{\mu}_{1,s} - \widehat{\mu}_{2,s} \le \sqrt{\frac{8\log^+(T_1/(2s))}{s}}\bigg)\\
&= \sum_{s=1}^{T/2} \mathbb{P}\bigg(\frac{\sum_{i=1}^{s} Z_i}{s} \le \sqrt{\frac{4}{s}\log^+\Big(\frac{T_1}{2s}\Big)} - \frac{\Delta}{\sqrt{2}}\bigg)\\
&\le \sum_{s=1}^{T} \mathbb{P}\bigg(-\frac{\sum_{i=1}^{s} Z_i}{s} + \sqrt{\frac{4}{s}\log^+\Big(\frac{T_1/2}{s}\Big)} \ge \frac{\Delta}{\sqrt{2}}\bigg)\\
&\le 1 + \frac{8\log^+(T_1\Delta^2/4)}{\Delta^2} + \frac{6}{\Delta^2} + \frac{2\sqrt{8\pi\log^+(T_1\Delta^2/4)}}{\Delta^2},
\end{aligned} \tag{C.30}$$

where the equality is by the definition of $\sum_{i=1}^{s} Z_i/s = \sum_{i=1}^{s}(X_i - Y_i - \Delta)/(\sqrt{2}s) = (\widehat{\mu}_{1,s} - \widehat{\mu}_{2,s} - \Delta)/\sqrt{2}$, and the last inequality is due to the first statement of Lemma C.5 since $-Z_i$ are 1-subgaussian variables as well.

Let

$$\epsilon_T = \sqrt{2\log(T\Delta^2)/(T_1\Delta^2)}. \tag{C.31}$$

Since we focus on the asymptotic optimality ($T \to \infty$) and $T_1 = \log^2 T$, we assume

$$\epsilon_T \in (0, 1/2) \qquad \textit{and} \qquad T\Delta^2 \ge 16e^3. \tag{C.32}$$

Let E be the event $\mu' \in [\mu_{1'} - \epsilon_T\Delta, \mu_{1'} + \epsilon_T\Delta]$. Applying Lemma C.2 and union bound, $\mathbb{P}(E) \ge 1 - 2/(T\Delta^2)$. Similar to (C.7), we have

$$\mathbb{E}[\tau_2] \le \mathbb{E}[\tau_2\,\mathbb{1}(E, 1' = 1)] + \mathbb{E}[\tau_2\,\mathbb{1}(E, 1' = 2)] + 2/\Delta^2. \tag{C.33}$$

To bound $\mathbb{E}[\tau_2\,\mathbb{1}(E, 1' = 1)]$, we assume event E holds and the chosen arm $1'$ is the best arm, i.e., $1' = 1$. Let $E_1 = \{E, 1' = 1\}$. Let $\Delta' = \mu' - \mathbb{E}[Y_{i+\tau_1}]$. Then conditioned on E_1, $\Delta' \in [(1-\epsilon_T)\Delta, (1+\epsilon_T)\Delta]$. Since $\epsilon_T \in (0, 1/2)$ and $T\Delta^2 \ge 16e^3$, we have that conditioned on E_1, $T(\Delta')^2 \ge (1-\epsilon_T)^2 T\Delta^2 \ge 4e^3$. Let $W_i = \mu' - Y_{i+\tau_1} - \Delta'$. Then $-W_i$ is 1-subgaussian random variable. By the stopping rule of *Stage III* in Algorithm 2, it holds that

$$\begin{aligned}
\mathbb{E}[\tau_2 \mid E_1] &\le \sum_{t_2=1}^{T} \mathbb{P}(\tau_2 \ge t_2 \mid E_1)\\
&= \sum_{t_2=1}^{T} \mathbb{P}\bigg(\mu' - \theta_{2',t_2} \le \sqrt{\frac{2}{t_2}\log\Big(\frac{T}{t_2}\Big(\log^2\frac{T}{t_2} + 1\Big)\Big)}\ \bigg|\ E_1\bigg)\\
&= \sum_{t_2=1}^{T} \mathbb{P}\bigg(-\frac{\sum_{i=1}^{t_2} W_i}{t_2} + \sqrt{\frac{2}{t_2}\log\Big(\frac{T}{t_2}\Big(\log^2\frac{T}{t_2} + 1\Big)\Big)} \ge \Delta'\ \bigg|\ E_1\bigg)
\end{aligned}$$

$$\leq 1 + \frac{3 + 2\log(4T\Delta^2(\log^2(4T\Delta^2)+1)) + \sqrt{4\pi \log(4T\Delta^2(\log^2(4T\Delta^2)+1))}}{(1-\epsilon_T)^2\Delta^2}. \tag{C.34}$$

where the last inequality is due to the second statement of Lemma C.5 and $-W_i$ are 1-subGuassian. Let $E_2 = \{E, 1' = 2\}$, using the same argument, we can derive same bound as in (C.34) for $\mathbb{E}[\tau_2 \mid E_2]$. Then We have

$$\begin{aligned}\Delta\mathbb{E}[\tau_2] &\leq \Delta\mathbb{E}[\tau_2 \mathbb{1}(E_1)] + \Delta\mathbb{E}[\tau_2 \mathbb{1}(E_2)] + \frac{2}{\Delta} \\ &\leq \Delta + \frac{2}{\Delta} + \frac{3 + 2\log(4T\Delta^2(\log^2(4T\Delta^2)+1)) + \sqrt{4\pi \log(4T\Delta^2(\log^2(4T\Delta^2)+1))}}{(1-\epsilon_T)^2\Delta}.\end{aligned} \tag{C.35}$$

Bounding term I_3: $\mathbb{P}(\tau_2 < T, a = 2)$ is the joint probability between the event that the chosen arm after *Stage III* is arm 2 and the event that the following stopping condition will be satisfied in *Stage III*:

$$|\mu' - \theta_{2',t_2}| < \sqrt{2/t_2 \log\big(T/t_2\big(\log^2(T/t_2)+1\big)\big)}. \tag{C.36}$$

Similar to (C.33),

$$I_3 \leq \Delta T\mathbb{P}[\tau_2 < T, a = 2 \mid E_1]\mathbb{P}[E_1] + \Delta T\mathbb{P}[\tau_2 < T, a = 2 \mid E_2]\mathbb{P}[E_2] + \frac{2}{\Delta}. \tag{C.37}$$

Again, we first assume E_1 holds. By definition, we have that conditioned on E_1, $\sum_i^s W_i/s = \mu' - \theta_{2',s} - \Delta'$ and W_i is 1-subgaussian with zero mean. Recall that we have $T(\Delta')^2 \geq 4e^3$. By the third statement of Lemma C.5, we have

$$\begin{aligned}\mathbb{P}(\tau_2 < T, a = 2 \mid E_1) &\leq \mathbb{P}\bigg(\exists t_2 \geq 1, \mu' - \theta_{2',t_2} + \sqrt{\frac{2}{t_2}\log\Big(\frac{T}{t_2}\Big(\log^2\frac{T}{t_2}+1\Big)\Big)} \leq 0 \,\bigg|\, E_1\bigg) \\ &\leq \mathbb{P}\bigg(\exists t_2 \geq 1, \mu' - \theta_{2',t_2} - \Delta' + \Delta' + \sqrt{\frac{2}{t_2}\log\Big(\frac{T}{t_2}\Big(\log^2\frac{T}{t_2}+1\Big)\Big)} \leq 0 \,\bigg|\, E_1\bigg) \\ &\leq \frac{4(16e^2+1)}{T(1-\epsilon_T)^2\Delta^2}.\end{aligned} \tag{C.38}$$

When E_2 holds, the proof is similar to the previous one. In particular, we only need to change the notations to $\Delta' = \mathbb{E}[X_{i+\tau_1}] - \mu'$, which satisfies conditioned on E_2, $\Delta' \in [(1-\epsilon_T)\Delta, (1+\epsilon_T)\Delta]$. Hence, we can derive same bound as (C.38) for term $\mathbb{P}(\tau_2 < T, a = 2 \mid E_2)$.
Therefore,

$$I_3 = \Delta T\mathbb{P}(\tau_2 < T, a = 2) \leq \frac{2}{\Delta} + \frac{4(16e^2+1)}{(1-\epsilon_T)^2\Delta}. \tag{C.39}$$

Completing the proof: Therefore, substituting (C.29), (C.30), (C.35) and (C.39) into (C.28), we have

$$R_\mu(T) \leq 2\Delta + \frac{40 + 8\log^+(T_1\Delta^2/4) + 2\sqrt{8\pi\log^+(T_1\Delta^2)}}{\Delta} \tag{C.40}$$

$$+\frac{4(16e^2+2)+2\log(4T\Delta^2(\log^2(4T\Delta^2)+1))+\sqrt{4\pi\log(4T\Delta^2(\log^2(4T\Delta^2)+1))}}{(1-\epsilon_T)^2\Delta}. \tag{C.41}$$

Recall that $\epsilon_T^2 = 2\log(T\Delta^2)/(T_1\Delta^2)$. Let $T_1 = \log^2 T$. When $T \to \infty$, we have $\epsilon_T \to 0$, and hence $\lim_{T\to\infty} R_\mu(T)/T = 2/\Delta$. ■

C.3. Proof of the Regret Bound of Algorithm 3

In this section, we provide the proof of Theorem 2.3. It will mostly follow the proof framework in Section C.2 for Theorem 2.2. Recall that in the proof of Theorem 2.2, we used the concentration inequalities in Lemma C.5 to upper bound τ_2, which is the total number of times that the suboptimal arm $2'$ is pulled in *Stage III* of Algorithm 2. Now in Line 4 of Algorithm 3, we added the extra stopping time $\log^2 T$ to *Stage III*. Therefore, Lemma C.5 is no longer directly applicable here. Instead, we need the following refined concentration lemma.

Lemma C.6 *Let $\delta \in (0, 2/\log^4 T)$ and $M_1, M_2, \ldots, M_n$ be 1-subgaussian random variables with zero means. Denote $\widehat{\mu}_n = \sum_{s=1}^n M_s/n$. Then the following inequality holds:*

$$\mathbb{P}\Bigg(\exists s \le \log^2 T : \widehat{\mu}_s + \sqrt{\frac{2}{s}\log\bigg(\frac{eT}{s}\bigg(\log^2\frac{T}{s}+1\bigg)\bigg)} - \delta \le 0\Bigg) \le \frac{16e^2\log T}{T}. \tag{C.42}$$

Proof [Proof of Theorem 2.3] For the sake of simplicity, we use the same notation that used in Theorem 2.2. Similar to (C.4), the regret of Algorithm 3 can be decomposed as follows

$$\begin{aligned} R_\mu(T) \le\ & \underbrace{\Delta T_1 \mathbb{P}(\tau_1 < T_1, 1' = 2)}_{I_1} + \underbrace{\Delta\mathbb{E}[\tau_1] + \Delta\mathbb{E}[\tau_2]}_{I_2} + \underbrace{\Delta T\mathbb{P}(\tau_2 < \log^2 T, a = 2)}_{I_3} \\ & + \underbrace{\mathbb{P}(\tau_2 = \log^2 T)R(IV \mid \tau_2 = \log^2 T)}_{I_4}, \end{aligned} \tag{C.43}$$

where terms I_1, I_2 and I_3 are the same as or slightly different from that in (C.4), and term I_4 is a new regret caused by Lines 14-19 of Algorithm 3, where $\tau_2 = \log^2 T$ and $R(IV \mid \tau_2 = \log^2 T)$ represents the regret of Lines 14-19 in *Stage IV*.

Proof of Asymptotic Optimality: The proof of the asymptotic optimality is almost the same as that in Section C.2. Recall the definition in (C.31) that $\epsilon_T = \sqrt{2\log(T\Delta^2)/(T_1\Delta^2)}$. To derive the asymptotic regret bound, since we consider the case that $T \to \infty$, we can trivially assume $\epsilon_T \in (0, 1/2)$ and $T\Delta^2 \ge 16e^3$. Note that *Stage I* and *Stage II* of Algorithm 3 are exactly the same as that of Algorithm 2. Based on the proof in Section C.2, it is easy to obtain the following results.

$$\Delta T_1 \mathbb{P}(\tau_1 < T_1, 1' = 2) = O\bigg(\frac{1}{\Delta}\bigg), \tag{C.44}$$

$$\Delta\mathbb{E}[\tau_1] = O\bigg(\Delta + \frac{\log^+(T_1\Delta^2)}{\Delta}\bigg), \tag{C.45}$$

$$\Delta\mathbb{E}[\tau_2] \le \Delta + \frac{O(1) + 2\log(4e\cdot T\Delta^2(\log^2(4e\cdot T\Delta^2)+1))}{(1-\epsilon_T)^2\Delta}$$

$$+\frac{\sqrt{4\pi\log(4e\cdot T\Delta^2(\log^2(4e\cdot T\Delta^2)+1))}}{(1-\epsilon_T)^2\Delta}. \tag{C.46}$$

which are due to (C.29), (C.30) and (C.35) respectively.

For term I_3, $\mathbb{P}(\tau_2 < \log^2 T, a = 2)$ is the joint probability between the event that the chosen arm after *Stage III* is the suboptimal arm 2 and the event that the following stopping condition will be satisfied after less than $\log^2 T$ time steps executed in *Stage III*:

$$|\mu' - \theta_{2',t_2}| < \sqrt{2/t_2\log\big(eT/t_2\big(\log^2(T/t_2)+1\big)\big)}. \tag{C.47}$$

Recall the proof in Section C.2 and note that the above probability is smaller than that in Algorithm 2 due to the extra requirement $\tau_2 < T$. Therefore, by (C.39) we have

$$I_3 = \Delta T\mathbb{P}(\tau_2 < \log^2 T, a = 2) = O\bigg(\frac{1}{(1-\epsilon_T)^2\Delta}\bigg). \tag{C.48}$$

Now, we bound the new term I_4. Note that $\tau_2 = \log^2 T$ implies Lines 15-19 is performed in *Stage IV*. Let τ_3 be the number of pulls of each arm in Line 16. Then the regret in Lines 15-19 can be upper bounded as $R(IV \mid \tau_2 = \log^2 T) \le \Delta\mathbb{E}[\tau_3] + \Delta T\mathbb{P}(\tau_3 < T, a = 2)$. Similar to the proof in (C.29), we have

$$\mathbb{P}(\tau_3 < T, a = 2) \le \mathbb{P}\Bigg(\exists s \in \mathbb{N} : 2s \le T,\ p_{1,s} - p_{2,s} \le -\sqrt{\frac{8\log(T/s)}{s}}\Bigg) \le \frac{15}{T\Delta^2}. \tag{C.49}$$

Similar to the proof of (C.30), we have

$$\begin{aligned}\mathbb{E}[\tau_3] = \sum_{s=1}^{T}\mathbb{P}(\tau_3 \ge s) &\le \sum_{s=1}^{T}\mathbb{P}\Bigg(p_{1s} - p_{2s} \le \sqrt{\frac{8\log(T/s)}{s}}\Bigg)\\ &\le 1 + \frac{8\log(T\Delta^2)}{\Delta^2} + \frac{6}{\Delta^2} + \frac{2\sqrt{8\pi\log(T\Delta^2)}}{\Delta^2}.\end{aligned} \tag{C.50}$$

Therefore, the regret generated by Lines 15-19 is

$$R(IV \mid \tau_2 = \log^2 T) = O\bigg(\Delta + \frac{\log(T\Delta^2)}{\Delta}\bigg). \tag{C.51}$$

To obtain the final bound for term I_4, we need to calculate the probability $\mathbb{P}(\tau_2 = \log^2 T)$. Since

$$\mathbb{E}[\tau_2] = \mathbb{E}[\tau_2|\tau_2 = \log^2 T] + \mathbb{E}[\tau_2|\tau_2 < \log^2 T] \ge \log^2 T\mathbb{P}(\tau_2 = \log^2 T), \tag{C.52}$$

combining the above result with (C.46), we have

$$\mathbb{P}(\tau_2 = \log^2 T) = O\bigg(\frac{\log(T\Delta^2)}{\Delta^2\log^2 T}\bigg). \tag{C.53}$$

Combining (C.51) and (C.53) together, we have

$$\lim_{T\to\infty}\frac{I_4}{\log T} = \lim_{T\to\infty}\frac{\mathbb{P}(\tau_2 = \log^2 T)R(IV \mid \tau_2 = \log^2 T)}{\log T} = 0.$$

In conclusion, substituting the above results back into the regret decomposition in (C.43), we have $\lim_{T\to\infty} R_\mu(T)/\log T = 2/\Delta$, which proves the asymptotic optimality of Algorithm 3.

Proof of Minimax/Instance-Dependent Optimality: When $T\Delta^2 \le 16e^3$, we have $R_\mu(T) \le T\Delta = O(\sqrt{T})$ and $R_\mu(T) \le T\Delta = O(1/\Delta)$, which is trivially minimax/instance-dependant optimal. Hence, we assume $T\Delta^2 \ge 16e^3$ in the rest of the proof. Different from the previous proof, ϵ_T defined in (C.31) may not fall in the interval $(0, 1/2)$ now. In particular, when the gap Δ is very small, the estimation of $\mu_{1'}$ will not be sufficiently accurate such that $\mu' \in [\mu_{1'} - \epsilon_T\Delta, \mu_{1'} + \epsilon_T\Delta]$. To handle this scenario, we will consider the following two cases.

Case 1: $\Delta > 1/\log^4 T$**.** Actually, if the unknown gap Δ is larger than $1/\log^4 T$, the proofs in the previous part for the asymptotic optimality still holds. Note that $T_1 = \log^{10} T$, then $\epsilon_T = \sqrt{2\log(T\Delta^2)/T_1\Delta^2} \in (0, 1/2)$. By the same argument as in (C.44), (C.45), (C.46) and (C.48), we have $I_1 + I_2 + I_3 = O(\Delta + \log(T\Delta^2)/\Delta)$. Also by (C.51), we have $I_4 \le R(IV|\tau_2 = \log^2 T) = O(\Delta + \log(T\Delta^2)/\Delta)$. Thus substituting these terms back into the regret decomposition in (C.43) yields $R_\mu(T) = O(\Delta + \log(T\Delta^2)/\Delta) = O(\Delta + \sqrt{T})$.

Case 2: $\Delta < 1/\log^4 T$**.** In this case, term I_1 and $\Delta\mathbb{E}[\tau_1]$ can be still bounded in the same way as in (C.44), (C.45), which leads to $I_1 + \Delta\mathbb{E}[\tau_1] = O(1/\Delta + \log^+(T_1\Delta^2)/\Delta)$.

Now we bound terms $\mathbb{E}[\tau_2]$ and I_3. Recall that in the previous part for proving the asymptotic regret, the bounds of term $\mathbb{E}[\tau_2]$ in (C.46) and term I_3 in (C.48) are heavily based on the results in (C.35) and (C.39). However, the results in (C.35) and (C.39) only hold based on the assumption $\epsilon_T \in (0, 1/2)$, which is not true in the case $\Delta < 1/\log^4 T$. Hence, (C.46) and (C.48) are not applicable here. Now, we bound these terms without assuming $\epsilon_T \in (0, 1/2)$. For term $\Delta\mathbb{E}[\tau_2]$, since we pull $2'$ at most $\log^2 T$ times in Stage *III* of Algorithm 3, it can be trivially seen that $\Delta\mathbb{E}[\tau_2] \le \Delta \log^2 T \le 1$.

For term I_3, note that we have pulled arm $1'$ for $T_1 = \log^{10} T$ times after *Stage II*. Applying Lemma C.2, we obtain

$$\mathbb{P}(|\mu' - \mu_{1'}| \ge 1/\log^4 T) \le 2/T^{1/2 \log T} \le 1/T,$$

where μ' is the average reward for arm $1'$ at the end of *Stage II* and we used the fact that $T \ge e^3$. Define event $E' = \{|\mu' - \mu_{1'}| \le 1/\log^4 T\}$ and its complement as E'^c. We further have

$$\begin{aligned} I_3 &\le \Delta T\mathbb{P}(\tau_2 < \log^2 T, a = 2 \mid E') + \Delta T\mathbb{P}(E'^c) \\ &\le \Delta T\mathbb{P}(\tau_2 < \log^2 T \mid E') + \Delta. \end{aligned} \tag{C.54}$$

Conditioned on event E', we have $|\mu' - \mu_{2'}| \le |\mu' - \mu_{1'}| + |\mu_{1'} - \mu_{2'}| \le 2/\log^4 T$ since $|\mu_{1'} - \mu_{2'}| = \Delta < 1/\log^4 T$. Based on this observation, we have

$$\begin{aligned} &\mathbb{P}(\tau_2 < \log^2 T \mid E') \\ &\le \mathbb{P}\left(\exists t_2 \le \log^2 T, |\mu' - \theta_{2',t_2}| \ge \sqrt{\frac{2}{t_2}\log\left(\frac{eT}{t_2}\left(\log^2\frac{T}{t_2} + 1\right)\right)} \,\middle|\, E'\right) \\ &\le \mathbb{P}\left(\exists t_2 \le \log^2 T, -(\mu' - \theta_{2',t_2}) + \sqrt{\frac{2}{t_2}\log\left(\frac{eT}{t_2}\left(\log^2\frac{T}{t_2} + 1\right)\right)} \le 0 \,\middle|\, E'\right) \\ &\quad + \mathbb{P}\left(\exists t_2 \le \log^2 T, -(\mu' - \theta_{2',t_2}) - \sqrt{\frac{2}{t_2}\log\left(\frac{eT}{t_2}\left(\log^2\frac{T}{t_2} + 1\right)\right)} \ge 0 \,\middle|\, E'\right) \end{aligned}$$

$$
\begin{aligned}
&\le \mathbb{P}\bigg(\exists t_2 \le \log^2 T, (\mu' - \mu_{2'}) - (\mu' - \theta_{2',t_2}) - |\mu' - \mu_{2'}| + \sqrt{\frac{2}{t_2}\log\bigg(\frac{eT}{t_2}\bigg(\log^2\frac{T}{t_2}+1\bigg)\bigg)} \le 0 \,\bigg|\, E'\bigg) \\
&\quad + \mathbb{P}\bigg(\exists t_2 \le \log^2 T, (\mu' - \mu_{2'}) - (\mu' - \theta_{2',t_2}) + |\mu' - \mu_{2'}| - \sqrt{\frac{2}{t_2}\log\bigg(\frac{eT}{t_2}\bigg(\log^2\frac{T}{t_2}+1\bigg)\bigg)} \ge 0 \,\bigg|\, E'\bigg) \\
&\le \frac{32e^2\log T}{T},
\end{aligned}
$$

where the first inequality is due to the stopping rule of *Stage III* in Algorithm 3, the second inequality is due to the fact that $\{|x-y| \ge z\} \subset \{x-y \ge z\} \bigcup \{x-y \le -z\}$, the third inequality is due to the fact that $\mu' - \mu_{2'} - |\mu' - \mu_{2'}| \le 0$ and $\mu' - \mu_{2'} + |\mu' - \mu_{2'}| \ge 0$, and in the last inequality we apply Lemma C.6 with $\delta = |\mu' - \mu_{2'}|$. Therefore, substituting the above inequality back into (C.54), we have the following bound for term I_3:

$$
\begin{aligned}
I_3 &\le \Delta T \mathbb{P}(\tau_2 < \log^2 T | E') + \Delta \\
&\le \Delta T \frac{16e^2 \log T}{T} + \Delta \le 1,
\end{aligned}
$$

where the last inequality is due to $\Delta < 1/\log^4 T$ and the fact that $T \ge e^3$. For term I_4, conditioned on $\tau_2 = \log^2 T$, the regret of *Stage IV* (namely, term $R(IV \mid \tau_2 = \log^2 T)$) only depends on the data collected in Lines 15-19 of Algorithm 3, which is therefore the same as in (C.51). we have

$$
I_4 \le R(IV \mid \tau_2 = \log^2 T) = O\bigg(\Delta + \frac{\log(T\Delta^2)}{\Delta}\bigg).
$$

Hence, for case 2, the total regret $R_\mu(T) = O(\Delta + \log(T\Delta^2)/\Delta) = O(\Delta + \sqrt{T})$. ■

C.4. Proof of the Regret Bound of Algorithm 6

Now we provide the proof of the regret bound of the anytime version DETC algorithm.

Proof The regret of Algorithm 6 is caused by pulling the suboptimal arm 2, which gives rise to $R_\mu(T) = \sum_{t=1}^{T} \Delta \mathbb{E}[\mathbb{1}\{A_t = 2\}]$. We will consider two intermediate points $t = \sqrt{\log T}$ and $t = \log^2 T$. Then we can decompose the regret of Algorithm 6 as follows:

$$
R_\mu(T) = \underbrace{\sum_{t=1}^{\sqrt{\log T}} \Delta \mathbb{E}[\mathbb{1}\{A_t = 2\}]}_{I_1} + \underbrace{\sum_{t=\sqrt{\log T}}^{\log^2 T} \Delta \mathbb{E}[\mathbb{1}\{A_t = 2\}]}_{I_2} + \underbrace{\sum_{t=\log^2 T}^{T} \Delta \mathbb{E}[\mathbb{1}\{A_t = 2\}]}_{I_3}. \quad \text{(C.55)}
$$

In what follows, we will bound these terms separately.

Bounding term I_1: Since the horizon length in this part is only $\sqrt{\log T}$, we can directly upper bound it as $I_1 \le \Delta\sqrt{T}$.

Bounding term I_2: Since Algorithm 6 has multiple epochs, we will rewrite the regret in I_2 in an epoch-wise fashion. Specifically, without loss of generality, we assume that there are two integers

r_1 and r_2 such that $2^{r_1} = \sqrt{\log T}$ and $2^{r_2} = \log^2 T$ respectively. Denote $\tau_{2,r}$ to be the total number of pulls of arm 2 in the r-th epoch, $r = 1, 2, \ldots$. Then we can rewrite I_2 in the following way:

$$I_2 = \sum_{t=\sqrt{\log T}}^{\log^2 T} \Delta \mathbb{E}[\mathbb{1}\{A_t = 2\}] = \sum_{r=r_1}^{r_2} \Delta \mathbb{E}[\tau_{2,r}]. \tag{C.56}$$

Note that the chosen arm $1'$ may be different in different epochs. In order to make the presentation more precise, we use $1'(r)$ to denote the arm that is chosen by Line 3 in the r-th epoch of Algorithm 6. Also note that at the beginning of epoch r, the current time step of the algorithm is $t = 2^r$. Let $\epsilon_T = 1/\log\log T$ and define event

$$E = \left\{ \bigcap_{r=r_1, r_1+1, \ldots, r_2} \left\{ |\widehat{\mu}_{1'(r)}(2^r) - \mu_{1'(r)}| < \epsilon_T \Delta \right\} \right\}.$$

Event E essentially says that at the beginning of any epoch $r \in [r_1, r_2]$, the average reward of the chosen arm $1'(r)$ is always close to its mean reward within a margin $\epsilon_T \Delta$. The characterization of this event is the key to analyzing the number of suboptimal arms pulled in each epoch.

Now we commute the probability that event E happens. For any $r \in [r_1, r_2]$, at the beginning of the r-th epoch, we know that the algorithm has run for 2^r times steps. Recall the definition of $T_k(t)$, the number of times that arm $1'(r)$ is pulled is $T_{1'(r)}(2^r)$. Since arm $1'(r)$ is the arm that has been pulled for the most times so far, it must have been pulled for more than $2^{r-1} \geq 2^{r_1 - 1} = \sqrt{\log T}/2$ times, namely, $T_{1'(r)}(2^r) \geq \sqrt{\log T}/2$. By Lemma C.2, we have

$$\begin{aligned}
\mathbb{P}\big(|\widehat{\mu}_{1'}(2^r) - \mu_{1'}| \geq \epsilon_T \Delta\big) &\leq 2 \exp\left(- \frac{T_{1'(r)}(2^r) \epsilon_T^2 \Delta^2}{2} \right) \\
&\leq 2 \exp\left(- \frac{\sqrt{\log T} \epsilon_T^2 \Delta^2}{4} \right) \\
&\leq \frac{2}{\log^4 T},
\end{aligned} \tag{C.57}$$

where the last inequality holds due to $\epsilon_T = 1/\log\log T$ and when T is sufficiently large T such that

$$\frac{\sqrt{\log T}}{4 \log\log T} \geq \frac{4(\log\log T)^2}{\Delta^2}. \tag{C.58}$$

Let E^c be the complement of event E. Then it holds that

$$\begin{aligned}
\mathbb{P}(E^c) &= \mathbb{P}\left(\left\{ \bigcap_{r=r_1, r_1+1, \ldots, r_2} \left\{ |\widehat{\mu}_{1'}(2^r) - \mu_{1'}| < \epsilon_T \Delta \right\} \right\}^c \right) \\
&= \mathbb{P}\left(\bigcup_{r=r_1, r_1+1, \ldots, r_2} \left\{ |\widehat{\mu}_{1'}(2^r) - \mu_{1'}| \geq \epsilon_T \Delta \right\} \right) \\
&\leq \sum_{r=r_1}^{r_2} \mathbb{P}(|\widehat{\mu}_{1'}(2^r) - \mu_{1'}| \geq \epsilon_T \Delta)
\end{aligned}$$

$$\leq 1/\log^3 T, \tag{C.59}$$

where in the first inequality we applied the union bound over all epochs $r \in [r_1, r_2]$, and the last inequality is due to (C.57) and $r_2 = 2\log_2 \log T \leq \log T/2$ for sufficiently large T.

Based on the characterization of event E, we bound the summation of $\tau_{2,r}$ in (C.56) as follows:

$$\begin{aligned}
\sum_{r=r_1}^{r_2} \mathbb{E}[\tau_{2,r}] &\leq \sum_{r=r_1}^{r_2} \mathbb{E}[\tau_{2,r}|E]\mathbb{P}(E) + \sum_{r=1}^{r_2} 2^r \mathbb{P}(E^c)\\
&\leq \sum_{r=r_1}^{r_2} \mathbb{E}[\tau_{2,r}|E]\mathbb{P}(E) + \frac{2^{r_2+1}}{\log^3 T}\\
&\leq \sum_{r=r_1}^{r_2} \mathbb{E}[\tau_{2,r}|E]\mathbb{P}(E) + \frac{2}{\log T}.
\end{aligned} \tag{C.60}$$

where in the first inequality we used the fact the $\tau_{2,r}$ is at most 2^r in the r-th epoch, the second inequality is due to (C.59), and the last inequality is due to $2^{r_2} = \log^2 T$.

In the r-th epoch of Algorithm 6, $\tau_{2,r}$ is contributed by two part: the number of pulls of arm 2 in Line 5 and the number of pulls of arm 2 in Line 9. We denote them as c_r^+ and c_r^- respectively such that $\tau_{2,r} = c_r^+ + c_r^-$. In epoch $r \in [r_1, r_2]$, by the fact that $\mathbb{E}[x] = \sum_s \mathbb{P}(x > s)$ we have

$$\begin{aligned}
&\mathbb{E}[c_r^+|E]\mathbb{P}(E)\\
&= \sum_{s=1}^{T} \mathbb{P}(c_r^+ \geq s \mid E)\mathbb{P}(E)\\
&\leq \sum_{t=2^r}^{2^{r+1}} \mathbb{P}\left(\widehat{\mu}_1(t) - \widehat{\mu}_2(t) \leq \sqrt{\frac{2}{T_2(t)} \log\left(\frac{r\cdot 2^r}{T_2(t)}\left(\log^2\left(\frac{r\cdot 2^r}{T_2(t)}\right)+1\right)\right)} \,\middle|\, E\right)\mathbb{P}(E)\\
&\leq \sum_{t=2^r}^{2^{r+1}} \mathbb{P}\left(\mu_1 - \epsilon_T\Delta - \widehat{\mu}_2(t) \leq \sqrt{\frac{2}{T_2(t)} \log\left(\frac{r\cdot 2^r}{T_2(t)}\left(\log^2\left(\frac{r\cdot 2^r}{T_2(t)}\right)+1\right)\right)} \,\middle|\, E\right)\mathbb{P}(E)\\
&\leq \sum_{t=1}^{2^{r+1}} \mathbb{P}\left(\widehat{\mu}_2(t) - \mu_1 + \Delta + \sqrt{\frac{2}{T_2(t)} \log\big(\frac{r\cdot 2^r}{T_2(t)}\big(\log^2(\frac{r\cdot 2^r}{T_2(t)})+1\big)\big)} \geq (1-\epsilon_T)\Delta\right),
\end{aligned} \tag{C.61}$$

where in the first inequality, $c_r^+ > 0$ (arm 2 is pulled in Line 5) means the arm chosen in this epoch is arm $1'(r) = 2$ and the stopping condition in Line 4 of Algorithm 6 is satisfied, in the second inequality, we used the fact that conditioned on event E, it holds that $\widehat{\mu}_1(t) \geq \mu_1 - \epsilon_T\Delta$, and in the last inequality, we used the fact that $\mathbb{P}(x|y)\mathbb{P}(y) = \mathbb{P}(x,y) \leq \mathbb{P}(x)$ for any random variables x and y. Now note that $\widehat{\mu}_2(t) - \mu_1 + \Delta = \widehat{\mu}_2(t) - \mu_2$ is 1-subgaussian with zero mean. Applying the second statement of Lemma C.5 with $\delta = (1-\epsilon_T)\Delta$, we have

$$\begin{aligned}
\sum_{r=r_1}^{r_2} \mathbb{E}[c_r^+|E]\mathbb{P}(E) &= \sum_{r=r_1}^{r_2} O\left(\frac{\log(r\cdot 2^r\Delta^2)}{(1-\epsilon_T)^2\Delta^2}\right)\\
&= 2\log\log T \cdot O\left(\frac{\log(r_2\cdot 2^{r_2}\Delta^2)}{\Delta^2}\right)
\end{aligned}$$

$$= O(\sqrt{\log T}), \tag{C.62}$$

where the last equality is from the upper bound of $1/\Delta^2$ in (C.58) and the following upper bound of Δ^2:

$$\Delta \le \log T \qquad \text{and} \qquad \log\log T \ge 4, \tag{C.63}$$

which holds for sufficiently large T.

Now we bound $\mathbb{E}[c_r^-|E]\mathbb{P}(E)$. Note that when $c_r^- > 0$ (arm 2 is pulled in Line 9), we know that (1) the stopping condition in Line 4 of Algorithm 6 is violated by some $t \le 2^{r+1}$; and (2) arm $a(r) = 2$. Therefore, we have

$$\begin{aligned}\mathbb{E}[c_r^-|E]\mathbb{P}(E) &= \mathbb{E}[c_r^-|E, c_r^- > 0]\mathbb{P}(E)\mathbb{P}(c_r^- > 0|E)\\ &= \mathbb{E}[c_r^-|E, c_r^- > 0]\mathbb{P}(E)\big[\mathbb{P}(c_r^- > 0, 1' = 1 \mid E) + \mathbb{P}(c_r^- > 0, 1' = 2 \mid E)\big].\end{aligned} \tag{C.64}$$

For the first term in (C.64), similar to the proof in (C.61), we have

$$\begin{aligned}&\mathbb{E}[c_r^-|E, c_r^- > 0]\mathbb{P}(E)\mathbb{P}(c_r^- > 0, 1' = 1 \mid E)\\ &\le 2^r\mathbb{P}\Bigg(\exists t \le 2^{r+1} : \mu_1 - \epsilon_T\Delta - \widehat{\mu}_2(t) < -\sqrt{\frac{2}{T_2(t)}\log\bigg(\frac{r\cdot 2^r}{T_2(t)}\bigg(\log^2\bigg(\frac{r\cdot 2^r}{T_2(t)}\bigg)+1\bigg)\bigg)}\Bigg)\\ &\le 2^{r+2}O\bigg(\frac{1}{r\cdot 2^r\Delta^2}\bigg),\end{aligned} \tag{C.65}$$

where in the first inequality we used the fact that $c_r^- \le 2^r$ and $\widehat{\mu}_1(t) \ge \mu_1 - \epsilon_T\Delta$, and the second inequality is due to third statement of Lemma C.5. Using exactly the same argument, we have

$$\begin{aligned}&\mathbb{E}[c_r^-|E, c_r^- > 0]\mathbb{P}(E)\mathbb{P}(c_r^- > 0, 1' = 2 \mid E)\\ &\le 2^r\mathbb{P}\Bigg(\exists t \le 2^{r+1} : \mu_2 + \epsilon_T\Delta - \widehat{\mu}_1(t) > \sqrt{\frac{2}{T_2(t)}\log\bigg(\frac{r\cdot 2^r}{T_2(t)}\bigg(\log^2\bigg(\frac{r\cdot 2^r}{T_2(t)}\bigg)+1\bigg)\bigg)}\Bigg)\\ &\le 2^{r+2}O\bigg(\frac{1}{r\cdot 2^r\Delta^2}\bigg).\end{aligned} \tag{C.66}$$

Therefore, it holds that

$$\sum_{r=r_1}^{r_2}\mathbb{E}[c_r^-|E]\mathbb{P}(E) = \sum_{r=1}^{r_2}2^{r+1}O\bigg(\frac{1}{r\cdot 2^r\Delta^2}\bigg) = O\bigg(\frac{\log r_2}{r\Delta^2}\bigg) = O\bigg(\frac{1}{\Delta^2}\bigg). \tag{C.67}$$

Combining (C.60), (C.62) and (C.67) together, we have

$$\begin{aligned}I_2 &= \Delta\sum_{r=r_1}^{r_2}\mathbb{E}[\tau_{2,r}]\\ &= \Delta\sum_{r=r_1}^{r_2}\mathbb{E}[c_r^+|E]\mathbb{P}(E) + \Delta\sum_{r=r_1}^{r_2}\mathbb{E}[c_r^-|E]\mathbb{P}(E) + \frac{2\Delta}{\log T}\end{aligned}$$

$$= O\bigg(\sqrt{\log T} + \frac{1}{\Delta} + \frac{\Delta}{\log T}\bigg). \tag{C.68}$$

Bounding term I_3: We start with decomposing I_3 into two terms. The first term is the number of pulls of arm 2 at Line 5, i.e., $\sum_{r=r_2+1}^{\log_2 T} c_r^+$ and the second term is the number of pulls of arm 2 at Line 9, i.e., $\sum_{r=r_2+1}^{\log_2 T} c_r^-$. Therefore, we have

$$I_3 = \mathbb{E}\Bigg[\sum_{r=r_2+1}^{\log_2 T} c_r^-\Bigg] + \mathbb{E}\Bigg[\sum_{r=r_2+1}^{\log_2 T} c_r^+\Bigg]. \tag{C.69}$$

Define event

$$E' = \Bigg\{\bigcap_{r=r_2+1,\ldots,\log_2 T} \big\{|\widehat{\mu}_{1'(r)}(2^r) - \mu_{1'(r)}| < \epsilon_T\Delta\big\}\Bigg\}.$$

E' says that for epoch $r \geq r_2 + 1$, the average reward of the chosen arm $1'(r)$ is close to its mean reward within a margin $\epsilon_T\Delta$, which plays a similar role as E does. Now, we commute the probability that E' happens. Since arm $1'(r)$ is the arm that has been pulled for the most times so far, $1'(r)$ must have been pulled for more than $2^{r_2} \geq \log^2 T$. By Lemma C.2, we have

$$\begin{aligned}
\mathbb{P}\big(|\widehat{\mu}_{1'}(2^r) - \mu_{1'}| \geq \epsilon_T\Delta\big) &\leq 2\exp\bigg(-\frac{T_{1'(r)}(2^r)\epsilon_T^2\Delta^2}{2}\bigg)\\
&\leq 2\exp\bigg(-\frac{\log^2 T\epsilon_T^2\Delta^2}{2}\bigg)\\
&\leq \frac{2}{T^2},
\end{aligned} \tag{C.70}$$

where the last inequality holds due to $\epsilon_T = 1/\log\log T$ and when T is sufficiently large T such that

$$\frac{\log^2 T}{2\log T} \geq \frac{2(\log\log T)^2}{\Delta^2}. \tag{C.71}$$

Let E'^c be the complement of event E. Then it holds that

$$\begin{aligned}
\mathbb{P}(E'^c) &= \mathbb{P}\Bigg(\Bigg\{\bigcap_{r=r_2+1,\ldots,\log_2 T} \big\{|\widehat{\mu}_{1'}(2^r) - \mu_{1'}| < \epsilon_T\Delta\big\}\Bigg\}^c\Bigg)\\
&= \mathbb{P}\Bigg(\bigcup_{r=r_2+1,\ldots,\log_2 T} \big\{|\widehat{\mu}_{1'}(2^r) - \mu_{1'}| \geq \epsilon_T\Delta\big\}\Bigg)\\
&\leq \sum_{r=r_2+1}^{\log_2 T} \mathbb{P}(|\widehat{\mu}_{1'}(2^r) - \mu_{1'}| \geq \epsilon_T\Delta)\\
&\leq 1/T,
\end{aligned} \tag{C.72}$$

where in the first inequality we applied the union bound over all epochs $r \in [r_2 + 1, \log_2 T]$, and the last inequality is due to (C.70) and $\log_2 T \leq T/2$. Recall that in (C.59), $\mathbb{P}(E^c) = O(1/\log^3 T)$.

However, this error probability is not sufficient to support the algorithm run for T steps. Instead, when we have pulled arm $1'$ for $\log^2 T$ times, since $\mathbb{P}(E'^c) \le 1/T$, the regret caused by E'^c is at most Δ.

Based on the characterization of event E', we bound the summation of c_r^+ and c_r^- in (C.69) as follows:

$$\begin{aligned}\sum_{r=r_2+1}^{\log_2 T} \mathbb{E}[c_r^+] + \sum_{r=r_2+1}^{\log_2 T} \mathbb{E}[c_r^-] &\le \sum_{r=r_2+1}^{\log_2 T} \mathbb{E}[c_r^+ \mid E']\mathbb{P}(E') + \sum_{r=r_2+1}^{\log_2 T} \mathbb{E}[c_r^- \mid E']\mathbb{P}(E') + \sum_{r=r_2+1}^{\log_2 T} 2^r \mathbb{P}(E'^c) \\ &\le \sum_{r=r_2+1}^{\log_2 T} \mathbb{E}[c_r^+ \mid E']\mathbb{P}(E') + \sum_{r=r_2+1}^{\log_2 T} \mathbb{E}[c_r^- \mid E']\mathbb{P}(E') + 2, \qquad \text{(C.73)}\end{aligned}$$

where in the first inequality we used the fact the $c_r^+ + c_r^-$ is at most 2^r in the r-th epoch, the second inequality is due to (C.72). Now, we bound term $\sum_{r=r_2}^{\log_2 T} \mathbb{E}[c_r^+|E']\mathbb{P}(E')$. Using the previous results (C.61) of bounding $\mathbb{E}[c_r^+|E]\mathbb{P}(E)$,

$$\begin{aligned}&\sum_{r=r_2+1}^{\log_2 T} \mathbb{E}[c_r^+|E']\mathbb{P}(E') \\ &\le \sum_{r=r_2+1}^{\log_2 T} \sum_{t=2^r}^{2^{r+1}} \mathbb{P}\bigg(\widehat{\mu}_2(t) - \mu_1 + \Delta + \sqrt{\frac{2}{T_2(t)} \log\bigg(\frac{r\cdot 2^r}{T_2(t)}\bigg(\log^2(\frac{r\cdot 2^r}{T_2(t)}) + 1\bigg)\bigg)} \ge (1-\epsilon_T)\Delta\bigg) \\ &\le \sum_{t=1}^{T} \mathbb{P}\bigg(\widehat{\mu}_2(t) - \mu_1 + \Delta + \sqrt{\frac{2}{T_2(t)} \log\bigg(\frac{\log_2 T\cdot T}{T_2(t)}\bigg(\log^2(\frac{\log_2 T\cdot T}{T_2(t)}) + 1\bigg)\bigg)} \ge (1-\epsilon_T)\Delta\bigg) \\ &\le \frac{2\log(T\Delta^2\log T) + o(\log(T\Delta^2\log T))}{(1-\epsilon_T)^2\Delta^2}, \qquad \text{(C.74)}\end{aligned}$$

where the last inequality is due to the second statement of Lemma C.5. By (C.66), we have

$$\mathbb{E}[c_r^-|E', c_r^- > 0]\mathbb{P}(E')\mathbb{P}(c_r^- > 0, 1' = 2 \mid E') \le 2^{r+2} O\bigg(\frac{1}{r\cdot 2^r\Delta^2}\bigg). \qquad \text{(C.75)}$$

Therefore, we have

$$\sum_{r=r_2+1}^{\log_2 T} \mathbb{E}[c_r^-|E']\mathbb{P}(E') = \sum_{r=r_2+1}^{r_2} 2^{r+1} O\bigg(\frac{1}{r\cdot 2^r\Delta^2}\bigg) = O\bigg(\int_{x=1}^{\log T} \frac{1}{x\Delta^2}\bigg)\mathrm{d}x = O\bigg(\frac{\log\log T}{\Delta^2}\bigg). \qquad \text{(C.76)}$$

Combing (C.73) and (C.76) together, we have

$$I_3 = \frac{2\log(T\Delta^2\log T) + o(\log(T\Delta^2\log T))}{(1-\epsilon_T)^2\Delta^2} + O\bigg(\frac{\log\log T}{\Delta^2}\bigg) + O(1). \qquad \text{(C.77)}$$

Substituting (C.68) and (C.77) into (C.55), we have

$$\lim_{T\to\infty} \frac{R_\mu(T)}{\log T} = \frac{2}{\Delta},$$

which completes the proof. ■

C.5. Proof of the Regret Bound of Algorithm 7

In this section, we prove the regret bound of DETC for K-armed bandits.

Proof [Proof of Theorem C.1] Let T_i be the total number of pulls of arm i throughout the algorithm, $i \geq 2$. Since by definition the regret is $R_\mu(T) = \sum_i \mathbb{E}[T_i \Delta_i]$, it suffices to prove

$$\lim_{T\to\infty} \frac{\mathbb{E}[T_i]}{\log(T)} = \frac{2}{\Delta_i^2}. \tag{C.78}$$

Denote $\tau_{2,i}$ as the number of pulls of arm i in *Stage III* of Algorithm 7. Similar to (C.3) and (C.4), the term $\mathbb{E}[T_i]$ can be decomposed as follows

$$\begin{aligned}\mathbb{E}[T_i] \leq \sqrt{\log T} + \underbrace{\log^2 T\mathbb{P}(1' = i)}_{I_1} + \underbrace{\mathbb{E}[\tau_{2,i}]}_{I_2} + \underbrace{T\mathbb{P}(\widehat{\mu}_{1'} \geq \theta_{j',t_j}, \mathcal{F}_{\text{fail}} = 0, a = i)}_{I_3} \\ + \underbrace{\log^2 T\mathbb{P}(\mathcal{F}_{\text{fail}} = 1) + T\mathbb{P}(\mathcal{F}_{\text{fail}} = 1, a = i)}_{I_4},\end{aligned} \tag{C.79}$$

where the last term I_4 characterizes the failing probability of the first three stages and the ETC step in the last two lines of Algorithm 7.

Bounding term I_1: Let $\widehat{\mu}_{i,s}$ be the estimated reward of arm i after its s-th pull. Let $\tau_1 = \sqrt{\log T}$. Let X be the reward of arm 1 and Y^i be the reward of arm i for $i > 1$. Let $S_n^i = X_1 - Y_1^i + \cdots + X_n - Y_n^i$. After pulling arm 1 and arm i τ_1 times, using Lemma C.2, we get

$$\mathbb{P}(S_{\tau_1}^i/\tau_1 \leq \Delta_i - \epsilon) \leq \exp(-\tau_1\epsilon^2/4). \tag{C.80}$$

For sufficiently large T such that $T > K$ and for all i, it holds

$$\frac{\sqrt{\log T}}{\log K + 2\log\log T} \geq \frac{4}{\Delta_i^2}, \tag{C.81}$$

Setting $\epsilon = \Delta_i$ in (C.80), we have $\mathbb{P}(\widehat{\mu}_{1,\tau_1} \leq \widehat{\mu}_{i,\tau_1}) \leq 1/(K\log^2 T)$. Applying union bound, we have

$$\mathbb{P}(\widehat{\mu}_{1,\tau_1} \geq \max_i \widehat{\mu}_{i,\tau_1}) = \mathbb{P}(1' = 1) \geq 1 - \frac{1}{\log^2 T}, \tag{C.82}$$

which further implies $I_1 \leq 1$.

Bounding term I_2: Let $\epsilon_i = \sqrt{4\log(T\Delta_i^2)/((\log T)^2\Delta_i^2)}$. Applying Lemma C.2, we have

$$\mathbb{P}(\mu' \notin (\mu_{1'} - \epsilon_i\Delta_i, \mu_{1'} + \epsilon_i\Delta_i)) \leq 2/(T\Delta_i^2). \tag{C.83}$$

Similar to (D.3), we choose a large T such that for all $\Delta_i > 0$,

$$\sqrt{\frac{4\log(T\Delta_i^2)}{\Delta_i^2\log^2 T}} \leq \frac{1}{(\log T)^{\frac{1}{3}}}, \tag{C.84}$$

then $\epsilon_i \leq 1/(\log T)^{\frac{1}{3}}$. Let E be the event $\mu' \in (\mu_{1'} - \epsilon_i\Delta_i, \mu_{1'} + \epsilon_i\Delta_i)$. Let E_1 be the event $\{E, 1' = 1\}$. Note that $\Pr(1' = 1) \geq 1 - 1/\log^2 T$, $\Pr(E^c) \leq 2/(T\Delta_i^2)$ and $\tau_{2,i} \leq \log^2 T$, the term I_2 can be decomposed as

$$\mathbb{E}[\tau_{2,i}] = \mathbb{E}[\tau_{2,i}\mathbb{1}(1' = 1)] + \mathbb{E}[\tau_{2,i}\mathbb{1}(1' \neq 1)]$$

$$\begin{aligned}&\le \mathbb{E}[\tau_{2,i}\,\mathbb{1}(1' = 1)] + 1\\&\le \mathbb{E}[\tau_{2,i}\,\mathbb{1}(E_1)] + \mathbb{E}[\tau_{2,i}\,\mathbb{1}(E^c)] + 1\\&\le 1 + \frac{2}{\Delta_i} + \mathbb{E}[\tau_{2,i} \mid E_1].\end{aligned} \tag{C.85}$$

We can derive the same bound as $\mathbb{E}[\tau_2 \mid E_1]$ in (C.34) for $\mathbb{E}[\tau_{2,i} \mid E_1]$. We have

$$\begin{aligned}I_2 &= \mathbb{E}[\tau_{2,i} \mid E_1]\\&\le 1 + \frac{3 + 2\log(4T\Delta_i^2(\log^2(4T\Delta_i^2)+1)) + \sqrt{4\pi\log(4T\Delta_i^2(\log^2(4T\Delta_i^2)+1))}}{(1-\epsilon_i)^2\Delta_i^2}.\end{aligned} \tag{C.86}$$

Bounding term I_3: When $\mathcal{F}_{\text{fail}} = 0$, we can follow the same proof for bounding I_3 in (C.38). Therefore, we can obtain

$$I_3 \le \frac{2}{\Delta_i^2} + \frac{4(16e^2+1)}{(1-\epsilon_i)^2\Delta_i^2}. \tag{C.87}$$

Bounding term I_4: For term $\mathbb{P}(\mathcal{F}_{\text{fail}} = 1)$, similar to (C.85), we have

$$\begin{aligned}\mathbb{P}(\mathcal{F}_{\text{fail}} = 1) &= \mathbb{P}(\mathcal{F}_{\text{fail}} = 1 \mid 1' = 1)\Pr(1' = 1) + \mathbb{P}(\mathcal{F}_{\text{fail}} = 1 \mid 1' \ne 1)\Pr(1' \ne 1)\\&\le \mathbb{P}(\mathcal{F}_{\text{fail}} = 1 \mid 1' = 1) + \frac{1}{\log^2 T}\\&\le \mathbb{P}(\mathcal{F}_{\text{fail}} = 1 \mid E, 1' = 1)\Pr(E \mid 1' = 1) + \Pr(E^c \mid 1' = 1) + \frac{1}{\log^2 T}\\&\le \mathbb{P}(\mathcal{F}_{\text{fail}} = 1 \mid E_1) + \frac{2}{T\Delta_i^2} + \frac{1}{\log^2 T},\end{aligned} \tag{C.88}$$

where the first and third inequalities are due to the law of total probability, the second inequality is due to (C.82), and the last inequality is due to (C.83). Let $\Delta_i' = \mu' - \mathbb{E}[Y_1^i]$, $W_r = \mu' - Y_{r+\tau_1}^i - \Delta_i'$. We have that conditioned on E_1, $\sum_r^s W_r/s = \mu' - \theta_{2',s} - \Delta'$ and W_r is 1-subgaussian with zero mean. By the third statement of Lemma C.5, we have

$$\begin{aligned}\mathbb{P}(\mathcal{F}_{\text{fail}} = 1 \mid E_1) &\le \mathbb{P}\biggl(\exists t_i \ge 1, \mu' - \theta_{i',t_i} + \sqrt{\frac{2}{t_i}\log\Bigl(\frac{T}{t_i}\Bigl(\log^2\frac{T}{t_i}+1\Bigr)\Bigr)} \le 0 \,\bigg|\, E_1\biggr)\\&\le \frac{4(16e^2+1)}{T(1-\epsilon_i)^2\Delta_i^2}.\end{aligned} \tag{C.89}$$

For term $T\mathbb{P}(\mathcal{F}_{\text{fail}} = 1, a = i)$, we choose large enough T to ensure

$$\exp(-\Delta_i^2\log^2 T/4) \le \frac{1}{T}. \tag{C.90}$$

Then, following the similar argument in (D.15), we can obtain

$$\mathbb{P}(\mathcal{F}_{\text{fail}} = 1, a = i) \le \frac{1}{T}. \tag{C.91}$$

Therefore, substituting (C.88), (C.89) and (C.91) into the definition of I_4 in (C.79), we have

$$\begin{aligned} I_4 &= \log^2 T\mathbb{P}(\mathcal{F}_{\text{fail}} = 1) + T\mathbb{P}(\mathcal{F}_{\text{fail}} = 1, a = i) \\ &\leq 2 + \frac{2\log^2 T}{T\Delta_i^2} + \frac{4(16e^2+1)\log^2 T}{T(1-\epsilon_i)^2\Delta_i^2}. \end{aligned} \tag{C.92}$$

Completing the proof: we can choose a sufficiently large T such that all the conditions (C.81), (C.84), (C.90) are satisfied simultaneously. Substituting (C.92), (C.87), (C.86) and $I_1 \leq 1$ back into (C.79), we have

$$\mathbb{E}[T_i] \leq 4 + \frac{C + 2\log(4T\Delta_i^2(\log^2(4T\Delta_i^2)+1)) + \sqrt{4\pi\log(4T\Delta_i^2(\log^2(4T\Delta_i^2)+1))}}{(1-\epsilon_i)^2\Delta_i^2},$$

for all $i \geq 2$, where $C > 0$ is a universal constant. Note that for $T \to \infty$, $\epsilon_i \leq 1/(\log T)^{\frac{1}{3}}$. Hence we have $\lim_{T\to\infty} \mathbb{E}[T_i]/\log T = 2/\Delta_i^2$ and $\lim_{T\to\infty} R_\mu(T)/\log T = \sum_i 2/\Delta_i$. ■

Appendix D. Round Complexity of Batched DETC

In this section, we derive the round complexities of Algorithms 4 and 5 for batched bandit models. We will prove that Batched DETC still enjoys the asymptotic optimality. Note that in batched bandits, our focus is on the asymptotic regret bound and thus we assume that T is sufficiently large throughout the proofs in this section to simplify the presentation.

D.1. Proof of Theorem 3.1

We first prove the round complexity for Batched DETC (Algorithm 4) when the gap Δ is known.

Proof The analysis is very similar to that of Theorem 2.1 and thus we will use the same notations therein. Note that *Stage I* requires 1 round of queries since τ_1 is fixed. In addition, *Stage II* and *Stage IV* need 1 query at the beginning of stages respectively. Now it remains to calculate the total rounds for *Stage III*.

Recall that E is event $\mu' \in [\mu_{1'} - \epsilon_T\Delta, \mu_{1'} + \epsilon_T\Delta]$, $E_1 = \{E, 1' = 1\}$ and $E_2 = \{E, 1' = 2\}$. We first assume that E_1 holds. Let $x_i = i(2\sqrt{\log(T\Delta^2)} + 4)$ and $n_{x_i} = \tau_0 + x_i/(2(1-\epsilon_T)^2\Delta^2)$. For simplicity, assume $x_i, n_{x_i} \in \mathbb{N}^+$. From (C.12), we have

$$\begin{aligned} \mathbb{P}(\tau_2 > n_{x_i} \mid E_1) \leq \mathbb{P}\bigg(S_{n_{x_i}} \leq \frac{\log(T\Delta^2)}{2(1-\epsilon_T)\Delta} \,\bigg|\, E_1\bigg) &\leq \exp\bigg(-\frac{x_i^2}{4(\log(T\Delta^2)+x_i)}\bigg) \\ &\leq \exp\bigg(-\frac{x_i}{2\sqrt{\log(T\Delta^2)}+4}\bigg) \\ &\leq 2^{-i}. \end{aligned} \tag{D.1}$$

Thus, the expected number of rounds of queries needed in *Stage III* of Algorithm 4 is upper bounded by $\sum_{i=1}^{\infty} i/2^i = 2$. Similarly, if E_2 holds, we still have the expected number of rounds in *Stage III* is upper bounded by 2. Lastly, if E^c holds, we have $\mathbb{P}(E^c) \leq 2/(T\Delta^2)$. Note that the increment between consecutive test time points is $(2\sqrt{\log(T\Delta^2)} + 4)/(2(1-\epsilon_T)^2\Delta^2)$, thus the expected number of test time points is at most $T(1-\epsilon_T)^2\Delta^2/(\sqrt{\log(T\Delta^2)})$. Then the expected number of rounds for this case is bounded by $2(1-\epsilon_T)^2/(\sqrt{\log(T\Delta^2)})$. For $T\to\infty$, the expected number of rounds cost for this case is 0. To summarize, the round complexity of Algorithm 4 is $O(1)$.

Following the same proof in (C.13) and (C.14), it is easy to verify that $\mathbb{E}[\tau_2 \mid E_1] \le \tau_0 + (2\sqrt{\log(T\Delta^2)}+4)/((1-\epsilon_T)^2\Delta^2)$, which is no larger than the bound in (C.14). The bounds for other terms remain the same. Therefore, the batched version of Algorithm 4 is still asymptotically optimal, instance-dependent optimal and minimax optimal. ■

D.2. Proof of Theorem 3.3

Now we prove the round complexity and regret bound for Batched DETC (Algorithm 5) when the gap Δ is unknown.

Proof For the sake of simplicity, we use the same notations that are used in Theorem 2.2 and its proof. To compute the round complexity and regret of *Stage I*, we first compute the probability that $\tau_1 > 2i\sqrt{\log T}$. We assume T is large enough such that it satisfies

$$\sqrt{\log T} \ge 16\log^+(T_1\Delta^2/2)/\Delta^2, \tag{D.2}$$

where we recall that $T_1 = \log^2 T$. Let $s_i = 2i\sqrt{\log T}$ for $i=1,2,\ldots$ and $\gamma = 4\log^+(T_1\Delta^2/2)/\Delta^2$. From (D.2), it is easy to verify that $s_i \ge 32i/\Delta^2$, $\gamma/s_i \le 1/8$ and $\sqrt{4\log^+(T_1/2s_i)/s_i} \le \Delta\sqrt{\gamma/s_i}$. The stopping rule in *Stage I* implies

$$\begin{aligned}
\mathbb{P}(\tau_1 \ge s_i) &\le \mathbb{P}\left(\widehat{\mu}_{1,s_i} - \widehat{\mu}_{2,s_i} \le \sqrt{\frac{8}{s_i}\log^+\left(\frac{T_1}{2s_i}\right)}\right)\\
&= \mathbb{P}\left(\frac{\sum_{i=1}^{s_i} Z_i}{s_i} \le \sqrt{\frac{4}{s_i}\log^+\left(\frac{T_1}{2s_i}\right)} - \frac{\Delta}{\sqrt{2}}\right)\\
&\le \mathbb{P}\left(\frac{\sum_{i=1}^{s_i} Z_i}{s_i} \le \Delta\sqrt{\frac{\gamma}{s_i}} - \frac{\Delta}{\sqrt{2}}\right)\\
&\le \exp\left(-\frac{s_i\Delta^2}{2}\left(\frac{1}{\sqrt{2}} - \sqrt{\frac{\gamma}{s_i}}\right)^2\right)\\
&\le \exp(-i)\\
&\le 2^{-i},
\end{aligned}$$

where the third inequality follows from Lemma C.2 and the fourth inequality is due to the fact that $s_i \ge 32i/\Delta^2$ and $\gamma/s_i \le 1/8$. Hence by the choice of testing points in (3.2), the expected number of rounds needed in *Stage I* of Algorithm 5 is upper bounded by $\sum_{i=1}^{\infty} i/2^i \le 2$. The expectation of τ_1 is upper bounded by $\mathbb{E}[\tau_1] \le \sum_{i=1}^{\infty} 2i\sqrt{\log T}/2^i \le 4\sqrt{\log T}$, which matches the bound derived in (C.30).

Now we focus on bounding term $\Delta\mathbb{E}[\tau_2]$ and the round complexity in *Stage III*. Let $\epsilon'_T = \sqrt{2}\epsilon_T = \sqrt{4\log(T\Delta^2)/(T_1\Delta^2)}$. Let E be the event $\mu' \in [\mu_{1'} - \epsilon'_T\Delta, \mu_{1'} + \epsilon'_T\Delta]$. Applying Lemma C.2, we have $\mathbb{P}(E^c) \le 1/(T^2\Delta^4)$. Hence, the expected number of test time points contributed by case E^c is $O(1/(T\Delta^4))$ which goes to zero when $T\to\infty$. Similarly, we assume that E holds and the chosen arm $1' = 1$. Recall $E_1 = \{E, 1' = 1\}$. Recall that this condition also implies

$\Delta' \in [(1-\epsilon'_T)\Delta, (1+\epsilon'_T)\Delta]$, where $\epsilon'_T = \sqrt{\log(T\Delta^2)/(T_1\Delta^2)}$ and $T_1 = \log^2 T$. When T is large enough such that it satisfies

$$\sqrt{\frac{4\log(T\Delta^2)}{\Delta^2\log^2 T}} \le \frac{1}{(\log T)^{\frac{1}{3}}}, \tag{D.3}$$

we have $\epsilon'_T \le 1/(\log T)^{\frac{1}{3}}$. Furthermore, we can also choose a large T such that

$$\sqrt{\log T}(\Delta')^2 \ge 2(\log\log T)^2. \tag{D.4}$$

Applying Lemma C.2, we have

$$\begin{aligned}\mathbb{P}\Big(\mu_{2'} - \Delta'(\log T)^{-\frac{1}{4}} \le \theta_{2',N_1} \le \mu_{2'} + \Delta'(\log T)^{-\frac{1}{4}} \mid E_1\Big) &\ge 1 - 2\exp\bigg(-\frac{2\log T(\Delta')^2}{2\sqrt{\log T}\log\log T}\bigg)\\ &\ge 1 - \frac{2}{\log^2 T},\end{aligned} \tag{D.5}$$

where the last inequality follows by (D.4). This means that after the first round of *Stage III* in Algorithm 5, the average reward for arm $2'$ concentrates around the true value $\mu_{2'}$ with a high probability. Let E_3 be the event $\mu_{2'} - \Delta'/\sqrt[4]{\log T} \le \theta_{2',N_1} \le \mu_{2'} + \Delta'/\sqrt[4]{\log T}$. Recall that $E_1 = \{E, 1' = 1\}$ and $E_2 = \{E, 1' = 2\}$. Let $H_1 = \{E_1, E_3\}$ and $H_2 = \{E_2, E_3\}$. We have

$$\begin{aligned}\mathbb{E}[\tau_2] &\le \mathbb{E}[\tau_2 \mid E_1, E_3]\mathbb{P}[E_1, E_3] + \mathbb{E}[\tau_2 \mid E_2, E_3]\mathbb{P}[E_2, E_3] + \mathbb{E}[\tau_2 \mid E_3^c]\mathbb{P}[E_3^c] + \mathbb{E}[\tau_2 \mid E^c]\mathbb{P}[E^c]\\ &\le \mathbb{E}[\tau_2 \mid H_1]\mathbb{P}[H_1] + \mathbb{E}[\tau_2 \mid H_2]\mathbb{P}[H_2] + \mathbb{E}[\tau_2 \mid E_3^c]\mathbb{P}[E_3^c] + 2/(T\Delta^3)\end{aligned} \tag{D.6}$$

We first focus on term $\mathbb{E}[\tau_2 \mid H_1]$. We assume event H_1 holds. Define

$$\begin{aligned}s'_i &= \frac{2(1+1/\sqrt[4]{\log T})^2\log(T\log^3 T)}{\widehat{\Delta}^2} + \frac{i(1+1/\sqrt[4]{\log T})^2(\log T)^{\frac{2}{3}}}{\widehat{\Delta}^2},\\ \gamma' &= \frac{2\log\big(T(\Delta')^2[\log^2(T(\Delta')^2)+1]\big)}{(\Delta')^2},\end{aligned}$$

for $i = 1, 2, \ldots$. Recall the definition of test time points in (3.3), we know that the $(i+1)$-th test in *Stage III* happens at time step $t_2 = s'_i$. We choose a large enough T such that

$$\log^3 T \ge (\Delta')^2(\log^2(T(\Delta')^2)+1). \tag{D.7}$$

Let $\Delta' = \mu' - \mu_{2'}$. Hence conditioned on H_1, $\widehat{\Delta} = \mu' - \theta_{2',N_1} \in [(1-1/\sqrt[4]{\log T})\Delta', (1+1/\sqrt[4]{\log T})\Delta']$. Then we have that conditioned on H_1

$$\frac{2(1+1/\sqrt[4]{\log T})^2\log(T\log^3 T)}{\widehat{\Delta}^2} \ge \frac{2\log(T\log^3 T)}{(\Delta')^2} \ge \gamma', \tag{D.8}$$

where the last inequality is due to (D.7). On the other hand, we also have that conditioned on H_1

$$s'_i \ge \frac{2(1+1/\sqrt[4]{\log T})^2\log(T\log^3 T)}{\widehat{\Delta}^2} \ge \frac{2}{(\Delta')^2}. \tag{D.9}$$

Therefore, by the definition of γ', it holds that conditioned on H_1

$$\Delta'\sqrt{\frac{\gamma'}{s'_i}} = \sqrt{\frac{2}{s'_i}\log(T(\Delta')^2(\log^2(T(\Delta')^2)+1))} \ge \sqrt{\frac{2}{s'_i}\log\bigg(\frac{T}{s'_i}\bigg(\log^2\bigg(\frac{T}{s'_i}\bigg)+1\bigg)\bigg)}.$$

Recall the definition $W_i = \mu' - Y_{i+\tau_1} - \Delta'$ used in (C.34). From the stopping rule of *Stage III* in Algorithm 2, conditioned on H_1, we obtain

$$\begin{aligned}
\mathbb{P}(\tau_2 \ge s'_i \mid H_1) &\le \mathbb{P}\bigg(\mu' - \theta_{2',s'_i} \le \sqrt{\frac{2}{s'_i}\log\bigg(\frac{T}{s'_i}\bigg(\log^2\frac{T}{s'_i}+1\bigg)\bigg)}\ \bigg|\ H_1\bigg)\\
&= \mathbb{P}\bigg(\frac{\sum_{i=1}^{s'_i} W_i}{s'_i} + \Delta' \le \sqrt{\frac{2}{s'_i}\log\bigg(\frac{T}{s'_i}\bigg(\log^2\frac{T}{s'_i}+1\bigg)\bigg)}\ \bigg|\ H_1\bigg)\\
&\le \exp\bigg(-\frac{s'_i(\Delta')^2}{2}\bigg(1-\sqrt{\frac{\gamma'}{s'_i}}\bigg)^2\bigg)\\
&= \exp\bigg(-\frac{(\Delta')^2}{2}(\sqrt{s'_i}-\sqrt{\gamma'})^2\bigg)\\
&= \exp\bigg(-\frac{(\Delta')^2}{2}\bigg(\frac{s'_i-\gamma'}{\sqrt{s'_i}+\sqrt{\gamma'}}\bigg)^2\bigg)\\
&\le \exp\bigg(-\frac{i^2(\log T)^{4/3}}{8s'_i(\Delta')^2}\bigg),
\end{aligned} \tag{D.10}$$

where the second inequality from Lemma C.2 and in the last inequality we used the fact that $s'_i - \gamma' \ge i(1+1/\sqrt[4]{\log T})^2(\log T)^{\frac{2}{3}}/(\widehat{\Delta}^2) \ge i(\log T)^{\frac{2}{3}}/(\Delta')^2$ by (D.8). Choose sufficiently large T to ensure

$$(\log T)^{\frac{4}{3}} \ge 8s'_i(\Delta')^2. \tag{D.11}$$

Substituting (D.11) back into (D.10) yields $\mathbb{P}(\tau_2 \ge s'_i \mid H_1) \le 1/2^i$. Similarly, $\mathbb{P}(\tau_2 \ge s'_i \mid H_2) \le 1/2^i$, Thus conditioned on H_1 (or H_2), the expected rounds used in *Stage III* of Algorithm 2 is upper bounded by $\sum_{i=1}^{\infty} i/2^i \le 2$. Recall that from (D.3), $\epsilon'_T \le 1/(\log T)^{\frac{1}{3}}$. Conditional on H_1, the expectation of τ_2 is upper bounded by

$$\begin{aligned}
\mathbb{E}[\tau_2 \mid H_1] &\le s'_1 + \sum_{i=2}[(s'_i - s'_1)\mathbb{P}(\tau_2 \ge s'_i \mid H_1)]\\
&\le \frac{2(1+1/(\log T)^{\frac{1}{4}})^2\log(T\log^3 T)}{\widehat{\Delta}^2} + \frac{2(1+1/\sqrt[4]{\log T})^2(\log T)^{\frac{2}{3}}}{\widehat{\Delta}^2}\\
&\le \frac{2(1+1/(\log T)^{\frac{1}{4}})^2\log(T\log^3 T) + 2(1+1/(\log T)^{\frac{1}{4}})^2(\log T)^{\frac{2}{3}}}{(1-1/(\log T)^{\frac{1}{3}})^2(1-1/(\log T)^{\frac{1}{4}})^2\Delta^2},
\end{aligned} \tag{D.12}$$

where the last inequality is due to $\Delta' \in [(1-\epsilon'_T)\Delta, (1+\epsilon'_T)\Delta]$. Similarly, we can derive same bound as in (D.12) for $\mathbb{E}[\tau_2 \mid H_2]$.

For the case E_3^c. Note that $\tau_2 \le \log^2 T$ and we have $\mathbb{P}(E_3^c) \le 2/\log^2 T$ by (D.5). Therefore $\mathbb{E}[\tau_2 \mid E_3^c]$ can be upper bounded by 2, which is dominated by (D.12). Since $\tau_2 \le \log^2 T$, conditioned on E_3^c, the expected rounds is upper bounded by $\mathbb{P}(E_3^c)\cdot\log^2 T \le 2$. To summarize, we have proved that conditioned on H_1 (or H_2, or E^c, or E_3^c), the expected rounds cost is $O(1)$. Therefore, the expected rounds cost of *Stage III* is $O(1)$.

Note that the above analysis does not change the regret incurred in *Stage III*. A slight difference of this proof from that of Theorem 2.2 arises when we terminate *Stage III* with $t_2 = \log^2 T$. The term I_3 can be written as

$$I_3 = \Delta T \mathbb{P}(\tau_2 = \log^2 T, a = 2) + \Delta T \mathbb{P}(\tau_2 < \log^2 T, a = 2), \tag{D.13}$$

We can derive same bound as (C.39) for term $\Delta T \mathbb{P}(\tau_2 < \log^2 T, a = 2)$. Now, we focus on term $\Delta T \mathbb{P}(\tau_2 = \log^2 T, a = 2)$. For this case, we have tested $\log^2 T$ samples for both arm 1 and 2. Let $G_0 = 0$ and $G_n = (X_1 - Y_{1+\tau_1}) + \cdots + (X_n - Y_{n+\tau_1})$ for every $n \geq 1$. Then $X_i - Y_{i+\tau_1} - \Delta$ is a $\sqrt{2}$-subgaussian random variable. Applying Lemma C.2 with $\epsilon = \Delta$ yields

$$\mathbb{P}\bigg(\frac{G_{\tau_2}}{\tau_2} \leq 0\bigg) \leq \exp\bigg(-\frac{\tau_2 \Delta^2}{4}\bigg).$$

Conditioned on $\tau_2 = \log^2 T$, we further obtain $\mathbb{P}(a = 2) = \mathbb{P}(G_{\tau_2} \leq 0) \leq \exp(-\Delta^2 \log^2 T/4) \leq 1/T$, where in the last inequality we again choose large enough T to ensure

$$\exp(-\Delta^2 \log^2 T/4) \leq \frac{1}{T}. \tag{D.14}$$

Therefore, we have proved that conditional on $\tau_2 = \log^2 T$,

$$\mathbb{P}(a = 2) \leq \frac{1}{T}. \tag{D.15}$$

Hence, $\Delta T \mathbb{P}(\tau_2 = \log^2 T, a = 2) \leq 1/\Delta$.

To summarize, we can choose a sufficiently large T such that all the conditions (D.2), (D.3), (D.4), (D.7), (D.11) and (D.14) are satisfied simultaneously. Then the round complexity of Algorithm 2 is $O(1)$. For the regret bound, since the only difference between Algorithm 5 and Algorithm 2 is the stopping rules of *Stage I* and *Stage III*, we only need to combine the regret for terms (D.12) and (D.15) and the fact that $\Delta \mathbb{E}[\tau_1] \leq 4\Delta\sqrt{\log T}$ to obtain the total regret. Therefore, we have $\lim_{T\to\infty} R(T)/\log T = 2/\Delta$. ■

Appendix E. Proof of Concentration Lemmas

In this section, we provide the proof of the concentration lemma and the maximal inequality for subgaussian random variables.

E.1. Proof of Lemma C.3

Our proof relies on the following maximal inequality for supermartingales.

Lemma E.1 (Ville (1939)) *If (S_n) is a non-negative supermartingale, then for any $x > 0$,*

$$\mathbb{P}\bigg(\sup_{n\in\mathbb{N}} S_n > x\bigg) \leq \frac{\mathbb{E}[S_0]}{x}.$$

Proof [Proof of Lemma C.3] The proof follows from the same idea as the proof of Lemma 4 (Maximal Inequality) in Ménard and Garivier (2017). If $\widehat{\mu}_n > 0$, then (C.24) holds trivially. Otherwise, if event $\{\exists N \le n \le M, \widehat{\mu}_n + \gamma \le 0\}$ holds, then the following three inequalities also hold simultaneously:

$$\widehat{\mu}_n \le 0, \qquad -\gamma\widehat{\mu}_n - \frac{\gamma^2}{2} \ge \gamma^2 - \frac{\gamma^2}{2} = \frac{\gamma^2}{2}, \quad \text{and} \quad -\gamma n\widehat{\mu}_n - \frac{n\gamma^2}{2} \ge \frac{N\gamma^2}{2},$$

where the second inequality is due to $\widehat{\mu}_n \le -\gamma$ and the last is due to $n \ge N$. Therefore, we have

$$\begin{aligned}
\mathbb{P}(\exists N \le n \le M, \widehat{\mu}_n + \gamma \le 0) &\le \mathbb{P}\bigg(\exists N \le n \le M, -\gamma n\widehat{\mu}_n - \frac{n\gamma^2}{2} \ge \frac{N\gamma^2}{2}\bigg) \\
&= \mathbb{P}\bigg(\max_{N\le n\le M} \exp\bigg(-\gamma n\widehat{\mu}_n - \frac{n\gamma^2}{2}\bigg) \ge \exp\bigg(\frac{N\gamma^2}{2}\bigg)\bigg) \\
&\le \mathbb{P}\bigg(\max_{1\le n\le M} \exp\bigg(-\gamma n\widehat{\mu}_n - \frac{n\gamma^2}{2}\bigg) \ge \exp\bigg(\frac{N\gamma^2}{2}\bigg)\bigg) \\
&\le \frac{\mathbb{E}[\exp(-\gamma X_1 - \gamma^2/2)]}{\exp(N\gamma^2/2)} \\
&\le \exp\bigg(-\frac{N\gamma^2}{2}\bigg),
\end{aligned}$$

where the third inequality is from Ville's maximal inequality (Ville, 1939) for non-negative supermartingale and the fact that $S_n = \exp(-\gamma n\widehat{\mu}_n - n\gamma^2/2)$ is a non-negative supermartingale. To show S_n is a non-negative supermartingale, we have

$$\begin{aligned}
\mathbb{E}[\exp(-\gamma n\widehat{\mu}_n - n\gamma^2/2)|S_1, \ldots, S_{n-1}] &= S_{n-1}\mathbb{E}[\exp(-\gamma X_n)]\exp(-\gamma^2/2) \\
&\le S_{n-1}\exp(\gamma^2/2)\exp(-\gamma^2/2) \\
&\le S_{n-1},
\end{aligned}$$

where the first inequality is from the definition of 1-subgaussian random variables. This completes the proof. ■

E.2. Proof of Lemma C.4

Proof Let $Z_i = (X_i - Y_i - \Delta)/\sqrt{2}$. Then Z_s is a 1-subgaussian random variable with zero mean. Applying the standard peeling technique, we have

$$\begin{aligned}
&\mathbb{P}\bigg(\exists s \ge 1 : \widehat{\mu}_s + \sqrt{\frac{8\log^+(N/s)}{s}} \le 0\bigg) \\
&\le \mathbb{P}\bigg(\exists s \ge 1 : \frac{\sum_{i=1}^s Z_i}{s} + \sqrt{\frac{4\log^+(N/s)}{s}} + \frac{\Delta}{\sqrt{2}} \le 0\bigg) \\
&\le \frac{15}{N\Delta^2},
\end{aligned} \tag{E.1}$$

where the last inequality is from Lemma 9.3 of Lattimore and Szepesvári (2020). ■

E.3. Proof of Lemma C.5

To prove Lemma C.5, we also need the following technical lemma from Ménard and Garivier (2017).

Lemma E.2 *For all $\beta > 1$ we have*

$$\frac{1}{e^{\log(\beta)/\beta} - 1} \le 2\max\{\beta, \beta/(\beta-1)\}. \tag{E.2}$$

Proof [Proof of Lemma C.5] For the first statement, let $\gamma = 4\log^+(T_1\delta^2)/\delta^2$. Note that for $n \ge 1/\delta^2$, it holds that

$$\delta\sqrt{\frac{\gamma}{n}} = \sqrt{\frac{4}{n}\log^+(T_1\delta^2)} \ge \sqrt{\frac{4}{n}\log^+\Big(\frac{T_1}{n}\Big)}. \tag{E.3}$$

Let $\gamma' = \max\{\gamma, 1/\delta^2\}$. Therefore, we have

$$\begin{aligned}\sum_{n=1}^{T}\mathbb{P}\bigg(\widehat{\mu}_n + \sqrt{\frac{4}{n}\log^+\Big(\frac{T_1}{n}\Big)} \ge \delta\bigg) &\le \gamma' + \sum_{n=\lceil\gamma\rceil}^{T}\mathbb{P}\bigg(\widehat{\mu}_n \ge \delta\bigg(1 - \sqrt{\frac{\gamma'}{n}}\bigg)\bigg)\\
&\le \gamma' + \sum_{n=\lceil\gamma'\rceil}^{\infty}\exp\bigg(-\frac{\delta^2(\sqrt{n}-\sqrt{\gamma'})^2}{2}\bigg) \qquad\qquad (\text{E.4})\\
&\le \gamma' + 1 + \int_{\gamma'}^{\infty}\exp\bigg(-\frac{\delta^2(\sqrt{x}-\sqrt{\gamma'})^2}{2}\bigg)\mathrm{d}x\\
&\le \gamma' + 1 + \frac{2}{\delta}\int_0^{\infty}\Big(\frac{y}{\delta} + \sqrt{\gamma'}\Big)\exp(-y^2/2)\mathrm{d}y\\
&\le \gamma' + 1 + \frac{2}{\delta^2} + \frac{\sqrt{2\pi\gamma'}}{\delta}, \qquad\qquad (\text{E.5})\end{aligned}$$

where (E.4) is the result of Lemma C.2 and (E.5) is due to the fact that $\int_0^\infty y\exp(-y^2/2)\mathrm{d}y = 1$ and $\int_0^\infty \exp(-y^2/2)\mathrm{d}y = \sqrt{2\pi}/2$. (E.5) immediately implies the claim in the first statement:

$$\begin{aligned}\sum_{n=1}^{T}\mathbb{P}\bigg(\widehat{\mu}_n + \sqrt{\frac{4}{n}\log^+\Big(\frac{T_1}{n}\Big)} \ge \delta\bigg) \le& \gamma' + \sum_{n=\lceil\gamma'\rceil}^{T}\mathbb{P}\bigg(\widehat{\mu}_n \ge \delta\bigg(1 - \sqrt{\frac{\gamma'}{n}}\bigg)\bigg)\\
\le& \gamma' + 1 + \frac{2}{\delta^2} + \frac{\sqrt{2\pi\gamma'}}{\delta}. \qquad\qquad (\text{E.6})\end{aligned}$$

Plugging $\gamma' \le 4\log^+(T_1\delta^2)/\delta^2 + 1/\delta^2$ to above equation, we obtain

$$\sum_{n=1}^{T}\mathbb{P}\bigg(\widehat{\mu}_n + \sqrt{\frac{4}{n}\log^+\Big(\frac{T_1}{n}\Big)} \ge \delta\bigg) \le 1 + \frac{4\log^+(T_1\delta^2)}{\delta^2} + \frac{3}{\delta^2} + \frac{\sqrt{8\pi\log^+(T_1\delta^2)}}{\delta^2}. \tag{E.7}$$

For the second statement, its proof is similar to that of the first one. Let us define the following quantity:

$$\rho = \frac{2\log(T\delta^2(\log^2(T\delta^2)+1))}{\delta^2}. \tag{E.8}$$

Note that for all $n \geq 1/\delta^2$, it holds that

$$\delta\sqrt{\frac{\rho}{n}} = \sqrt{\frac{2\log(T\delta^2(\log^2(T\delta^2)+1))}{n}} \geq \sqrt{\frac{2}{n}\log\left(\frac{T}{n}\left(\log^2\frac{T}{n}+1\right)\right)}. \tag{E.9}$$

Using the same argument in (E.5) we can show that

$$\sum_{n=1}^{T}\mathbb{P}\left(\widehat{\mu}_n + \sqrt{\frac{2}{n}\log\left(\frac{T}{n}\left(\log^2\frac{T}{n}+1\right)\right)} \geq \delta\right) \leq 1 + \frac{2\log(T\delta^2(\log^2(T\delta^2)+1))}{\delta^2} + \frac{3}{\delta^2} + \frac{\sqrt{4\pi\log(T\delta^2(\log^2(T\delta^2)+1))}}{\delta^2}.$$

To prove the last statement, we borrow the idea from Ménard and Garivier (2017) for proving the regret of kl-UCB^{++}. Define $f(\delta) = 2/\delta^2\log(T\delta^2/4)$. Then we can decompose the event $\{\exists s : s \leq T\}$ into two cases: $\{\exists s : s \leq f(\delta)\}$ and $\{\exists s : f(\delta) \leq s \leq T\}$.

$$\begin{aligned}&\mathbb{P}\left(\exists s \leq T : \widehat{\mu}_s + \sqrt{\frac{2}{s}\log\left(\frac{T}{s}\left(\log^2\frac{T}{s}+1\right)\right)} + \delta \leq 0\right)\\ &\leq \underbrace{\mathbb{P}\left(\exists s \leq f(\delta) : \widehat{\mu}_s \leq -\sqrt{\frac{2}{s}\log\left(\frac{T}{s}\left(\log^2\frac{T}{s}+1\right)\right)}\right)}_{A_1} + \underbrace{\mathbb{P}(\exists s, f(\delta) \leq s \leq T : \widehat{\mu}_s \leq -\delta)}_{A_2}.\end{aligned} \tag{E.10}$$

Note that when $T\delta^2 \geq 4e^3$, $f(\delta) \geq 0$. Let $\beta > 1$ be a parameter that will be chosen later. Applying the peeling technique, we can bound term A_1 as follows.

$$A_1 \leq \sum_{\ell=0}^{\infty}\underbrace{\mathbb{P}\left(\exists s, \frac{f(\delta)}{\beta^{\ell+1}} \leq s \leq \frac{f(\delta)}{\beta^{\ell}} : \widehat{\mu}_s + \sqrt{\frac{2}{s}\log\left(\frac{T}{s}\left(\log^2\frac{T}{s}+1\right)\right)} \leq 0\right)}_{A_1^{\ell}}. \tag{E.11}$$

For each $\ell = 0, 1, \ldots$, define γ_l to be

$$\gamma_\ell = \frac{\beta^\ell}{f(\delta)}\log\left(\frac{T\beta^\ell}{2f(\delta)}\left(1+\log^2\frac{T}{2f(\delta)}\right)\right), \tag{E.12}$$

which by definition immediately implies

$$\sqrt{2\gamma_l} = \sqrt{\frac{2\beta^\ell}{f(\delta)}\log\left(\frac{T\beta^\ell}{2f(\delta)}\left(1+\log^2\frac{T}{2f(\delta)}\right)\right)} \leq \sqrt{\frac{2}{s}\log\left(\frac{T}{2s}\left(\log^2\frac{T}{s}\right)+1\right)},$$

where in the above inequality we used the fact that $s \leq f(\delta)/\beta^\ell$ and that $f(\delta) \geq s/2$ since $\beta > 1$. Therefore, we have

$$\mathbb{P}\left(\exists s, \frac{f(\delta)}{\beta^{\ell+1}} \leq s \leq \frac{f(\delta)}{\beta^{\ell}} : \widehat{\mu}_s + \sqrt{\frac{2}{s}\log\left(\frac{T}{s}\left(\log^2\frac{T}{s}+1\right)\right)} \leq 0\right)$$

$$
\begin{aligned}
&\leq \mathbb{P}\bigg(\exists \frac{f(\delta)}{\beta^{\ell+1}} \leq s \leq \frac{f(\delta)}{\beta^{\ell}} : \widehat{\mu}_s + \sqrt{2\gamma_\ell} \leq 0\bigg) \\
&\leq \exp\bigg(-\frac{f(\delta)}{\beta^{\ell+1}}\gamma_\ell\bigg) \\
&= e^{-\ell\log(\beta)/\beta - C/\beta},
\end{aligned} \tag{E.13}
$$

where the second inequality is by Doob's maximal inequality (Lemma C.3), the last equation is due to the definition of γ_ℓ, and the parameter C is defined to be

$$
C := \log\bigg(\frac{T}{2f(\delta)}\bigg(1 + \log^2 \frac{T}{2f(\delta)}\bigg)\bigg). \tag{E.14}
$$

Substituting (E.13) back into (E.11), we get

$$
A_1 \leq \sum_{\ell=0}^{\infty} e^{-\ell\log(\beta)/\beta - C/\beta} = \frac{e^{-C/\beta}}{1 - e^{-\log(\beta)/\beta}} \leq \frac{e^{1-C/\beta}}{e^{\log(\beta)/\beta} - 1} \leq 2e\max(\beta, \beta/(\beta-1))e^{-C/\beta},
$$

where the second inequality is due to $\log\beta \leq \beta$ and thus $e^{\log(\beta)/\beta} \leq e$, and the last inequality comes from Lemma E.2. Since $T\delta^2 \geq 4e^3$, we have $T/(2f(\delta)) = T\delta^2/(4\log(T\delta^2/4)) \geq \sqrt{T\delta^2/4} \geq e^{3/2}$, which further implies

$$
C = \log\bigg(\frac{T}{2f(\delta)}\bigg(1 + \log^2 \frac{T}{2f(\delta)}\bigg)\bigg) \geq \log\bigg(\frac{T}{2f(\delta)}\bigg) = \log\bigg(\frac{T\delta^2}{4\log(\frac{T\delta^2}{4})}\bigg) \geq 3/2. \tag{E.15}
$$

Now we choose $\beta := C/(C-1)$, so that $1 < \beta \leq 2C$ and $\beta/(\beta-1) = C$. Together with the definition of f, this choice immediately yields

$$
A_1 \leq 4eCe^{-C/\beta} = 4e^2Ce^{-C}. \tag{E.16}
$$

Note that

$$
\begin{aligned}
Ce^{-C} &= \bigg(\frac{T}{2f(\delta)}\bigg(1 + \log^2 \frac{T}{2f(\delta)}\bigg)\bigg)^{-1} \log\bigg(\frac{T}{2f(\delta)}\bigg(1 + \log^2 \frac{T}{2f(\delta)}\bigg)\bigg) \\
&\leq \frac{2f(\delta)}{T\log^2(T/(2f(\delta)))} \log\bigg(\frac{T}{2f(\delta)}\bigg(1 + \log^2 \frac{T}{2f(\delta)}\bigg)\bigg) \\
&\leq \frac{4f(\delta)}{T\log(T/(2f(\delta)))} \\
&= \frac{8\log(T\delta^2/4)}{T\delta^2\log([T\delta^2/4]/\log(T\delta^2/4))} \\
&\leq \frac{16}{T\delta^2},
\end{aligned} \tag{E.17}
$$

where in the second and the third inequalities, we used the fact that that for all $x \geq e^{3/2}$,

$$
\frac{\log(x(1 + \log^2 x))}{\log x} \leq 2 \qquad \text{and} \qquad \frac{\log x}{\log(x/\log x)} \leq 2. \tag{E.18}
$$

Therefore, we have proved so far $A_1 \leq 64e^2/(T\delta^2)$. For term A_2 in (E.10), we can again apply the maximal inequality in Lemma C.3 and obtain

$$A_2 = \mathbb{P}(\exists s, f(\delta) \leq s \leq T : \widehat{\mu}_s \leq -\delta) \leq e^{-\delta^2 f(\delta)/2} = \frac{4}{T\delta^2}. \tag{E.19}$$

Finally, combining the above results, we get

$$\mathbb{P}\bigg(\exists s \leq f(\delta), \widehat{\mu}_s + \sqrt{\frac{2}{s}\log\left(\frac{T}{s}\left(\log^2\frac{T}{s}+1\right)\right)} + \delta \leq 0\bigg) \leq \frac{4(16e^2+1)}{T\delta^2}. \tag{E.20}$$

This completes the proof. ■

E.4. Proof of Lemma C.6

Proof Recall $\delta \in (0, 2/\log^4 T)$. Note that for $s \leq \log^2 T$,

$$\begin{aligned}\sqrt{\frac{2}{s}\log\left(\frac{eT}{s}\left(\log^2\frac{T}{s}+1\right)\right)} - \delta &\geq \sqrt{\frac{2}{s}\left(1+\log\left(\frac{T}{s}\left(\log^2\frac{T}{s}+1\right)\right)\right)} - \frac{2}{\log^4 T}\\ &\geq \sqrt{\frac{2}{s}\log\left(\frac{T}{s}\left(\log^2\frac{T}{s}+1\right)\right)}.\end{aligned}$$

Let $a(s) = 2/s, b(s) = 2/s\log(T/s(\log^2(T/s)+1)$. The last inequality is equals to $\sqrt{a(s)+b(s)} - \sqrt{b(s)} \geq 2/\log^4 T$ for $s \leq \log^2 T$, which holds because (i):

$$\sqrt{a(s)+b(s)} - \sqrt{b(s)} = a(s)/(\sqrt{a(s)+b(s)} + \sqrt{b(s)});$$

(ii): for $s \leq \log^2 T$, then $a(s) \geq a(\log^2 T) = 2/\log^2 T$,

$$\sqrt{a(s)+b(s)} + \sqrt{b(s)} \leq \sqrt{a(1)+b(1)} + \sqrt{b(1)} < \log^2 T,$$

hence $a(s)/(\sqrt{a(s)+b(s)} + \sqrt{b(s)}) \geq 2/\log^4 T$. Now, we only need to prove

$$\mathbb{P}\bigg(\exists s \leq \log^2 T : \widehat{\mu}_s + \sqrt{\frac{2}{s}\log\left(\frac{T}{s}\left(\log^2\frac{T}{s}+1\right)\right)} \leq 0\bigg) \leq \frac{16e^2\log T}{T}. \tag{E.21}$$

The rest proof of Lemma C.6 is similar to the proof of Lemma C.5. Let A_1 be the r.h.s. (E.21) and $f = \log^2 T$. Then applying the peeling technique, we can bound A_1 as follows.

$$A_1 \leq \sum_{\ell=0}^{\infty}\mathbb{P}\bigg(\exists\frac{f}{\beta^{\ell+1}} \leq s \leq \frac{f}{\beta^\ell} : \widehat{\mu}_s + \sqrt{\frac{2}{s}\log\left(\frac{T}{s}\left(\log^2\frac{T}{s}+1\right)\right)} \leq 0\bigg).$$

Similar to (E.16), we have $A_1 \leq 2e\max(\beta, \beta/(\beta-1))e^{-C/\beta} \leq 4e^2Ce^{-C}$. Then (E.17) becomes

$$Ce^{-C} = \left(\frac{T}{2f}\left(1+\log^2\frac{T}{2f}\right)\right)^{-1}\log\left(\frac{T}{2f}\left(1+\log^2\frac{T}{2f}\right)\right)$$

$$\begin{aligned} &\leq \frac{4f}{T\log(T/(2f))} \\ &\leq \frac{4\log T}{T}, \end{aligned} \tag{E.22}$$

where the last inequality is due to $f = \log^2 T$. Similar to the proof in Lemma C.5, we have $A_1 \leq 16e^2 \log T/T$. Therefore, we have

$$\mathbb{P}\Bigg(\exists s \leq \log^2 T : \widehat{\mu}_s + \sqrt{\frac{2}{s}\log\bigg(\frac{eT}{s}\bigg(\log^2\frac{T}{s}+1\bigg)\bigg)} - \delta \leq 0\Bigg) \leq \frac{16e^2\log T}{T}.$$

This completes the proof. ■