Appendix A. Extension of DETC to the Unknown Gap Setting

In real world applications, the suboptimal gap Δ is often unknown. Thus, it is favorable to design an algorithm without the knowledge of Δ. In this section, we propose a variant of DETC for the unknown gap setting and further extend it to the batched bandit setting.

A.1. The Proposed DETC Algorithm

When the suboptimal gap Δ is unknown, the DETC algorithm is displayed in Algorithm 3. Similar to Algorithm 1, Algorithm 3 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 do not have access to 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 $|\widehat{\mu}_1(t)-\widehat{\mu}_2(t)| \geq \sqrt{16\log^+(T_1/t)/t}$, 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 choose 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*. Note that in both exploration stages, we do not need the information of the suboptimal gap Δ.

In the following theorem, we present the regret bound of Algorithm 3 and show that this regret is also asymptotically optimal and minimax optimal in this setting.

Theorem 4 *Let $\epsilon_T = \sqrt{2\log(T\Delta^2)/(T_1\Delta^2)}$. Suppose that $\epsilon_T \in (0,1/2)$ and $T\Delta^2 \geq 16e^3$, then*

$$R_\mu(T) \leq 2\Delta + \frac{38 + 8\log^+(T_1\Delta^2/4) + 2\sqrt{8\pi\log^+(T_1\Delta^2)}}{\Delta}$$
$$+ \frac{4(16e^2+2) + 2\log(4T\Delta^2(\log^2(4T\Delta^2)+1)) + \sqrt{4\pi\log(4T\Delta^2(\log^2(4T\Delta^2)+1))}}{(1-\epsilon_T)^2\Delta}.$$

Moreover,if we choose $T_1 = \log^2 T$, then $\lim_{T\to\infty} R_\mu(T)/\log T = 2/\Delta$.

Here we provide some comparison between existing algorithms and Algorithm 3. 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 4) separates the exploration and exploitation stages, which is much more practical in many real world applications such as clinic 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 A.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 our DETC (Algorithm 4) improves this regret bound to $2/\Delta$. Therefore, in order to break the $4/\Delta$ barrier in the asymptotic regret rate, our double exploration technique in Algorithm 3 is crucial.

Algorithm 3 Double Explore-then-Commit (DETC) for Unknown Gaps

```
input $T, T_1$
Initialization: $A_1 = 1$, $A_2 = 2$, $t \leftarrow 2$

Stage I: Explore all arms uniformly
while $|\ \widehat{\mu}_1(t) - \widehat{\mu}_2(t)\ | < \sqrt{16/t \log^+(T_1/t)}$ do
    Choose $A_{t+1} = 1$ and $A_{t+2} = 2$, $t \leftarrow t + 2$;
end while

Stage II: Commit the arm with the largest average reward
$1' \leftarrow \arg\max_i \widehat{\mu}_i(t)$;
while $t \le T_1$ do
    Choose $A_{t+1} = 1'$, $t \leftarrow t + 1$;
end while

Stage III: Explore the unchosen arm in Stage II
$\mu' \leftarrow \widehat{\mu}_{1'}(t)$, $2' \leftarrow \{1, 2\} \setminus 1'$, $t_2 \leftarrow 0$, $\theta_{2's}$ 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 $|\mu' - \theta_{2',t_2}| < \sqrt{2/t_2 \log\big(T/t_2\big(\log^2(T/t_2) + 1\big)\big)}$ do
    $A_{t+1} = 2'$, $t \leftarrow t + 1$, $t_2 \leftarrow t_2 + 1$;
end while

Stage IV: Commit the arm with the largest average reward after double exploration
$a \leftarrow 1' \mathbb{1}\{\widehat{\mu}_{1'}(t) \ge \theta_{2',t_2}\} + 2' \mathbb{1}\{\widehat{\mu}_{1'}(t) < \theta_{2',t_2}\}$;
while $t \le T$ do
    Play arm $a$, $t \leftarrow t + 1$.
end while
```

Similar to Theorem 1, one can verify that the regret in Theorem 4 is in the order of $O(\Delta + 1/\Delta \log(T\Delta^2)$ if we choose $T_1 = T$, which implies that Algorithm 2 is instance-dependent optimal and minimax optimal. However, in this case, we will only have the first two stages in Algorithm 3 and thus the DETC algorithm reduces to the single explore-then-commit algorithm in Garivier et al. (2016). It is an interesting open problem whether our DETC algorithm can achieve the improved asymptotic regret rate $2/\Delta$ without compromise on the instance-dependent/minimax optimality.

A.2. Batched DETC in the Unknown Gap Setting

In the unknown gap case, both the stopping rules of *Stage I* and *Stage III* in Algorithm 3 need to be modified. In what follows, we describe a variant of Algorithm 3 that only needs to check the results of pulls 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 2 of Algorithm 4 at the following time grid:

$$t \in \mathcal{T}_2 = \{2\sqrt{\log T}, 4\sqrt{\log T}, 6\sqrt{\log T}, \ldots\}. \tag{7}$$

In *Stage III*, we the condition in Line 10 of Algorithm 4 is only checked at the following time grid.

$$\begin{aligned} t_2 \in \mathcal{T}_2' = \big\{ & N_1, 2/\widehat{\Delta}^2 N_2 \log(T \log^3 T) + 1/\widehat{\Delta}^2 N_2 (\log T)^{\frac{2}{3}}, \\ & 2/\widehat{\Delta}^2 N_2 \log(T \log^3 T) + 2/\widehat{\Delta}^2 N_2 (\log T)^{\frac{2}{3}}, \\ & 2/\widehat{\Delta}^2 N_2 \log(T \log^3 T) + 3/\widehat{\Delta}^2 N_2 (\log T)^{\frac{2}{3}}, \cdots, \cdots, \log^2 T \big\}. \end{aligned} \tag{8}$$

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 3 is that we require $t_2 \leq T_1$. Thus we will terminate *Stage III* after at most $T_1 = \log^2 T$ pulls of arm $2'$. For the convenience of readers, we display the modified Algorithm 3 for batched bandits with unknown gaps in Algorithm 4.

Algorithm 4 Batched DETC in the Unknown Gap Setting

```
input $T$, $T_1$, $\mathcal{T}_2$ defined in (7), and $\mathcal{T}_2'$ defined in (8)
 1: Initialization: $A_1 = 1$, $A_2 = 2$, $t \leftarrow 2$
    ------------------------------------------------------------
    Stage I: Explore all arms uniformly
 2: while true do
 3:     if $t \in \mathcal{T}_2$ then
 4:         if $| \widehat{\mu}_1(t) - \widehat{\mu}_2(t) | \geq \sqrt{16/t \log^+(T_1/t)}$ then
 5:             break
 6:         end if
 7:     end if
 8:     Choose $A_{t+1} = 1$ and $A_{t+2} = 2$, $t \leftarrow t + 2$;
 9: end while
10: Stage II
11: ⋮
12: ------------------------------------------------------------
    Stage III: Explore the unchosen arm in Stage II
13: $\mu' \leftarrow \widehat{\mu}_{1'}(t)$, $2' \leftarrow \{1, 2\} \setminus 1'$, $t_2 \leftarrow 0$, $\theta_{2's}$ is the recalculated average reward of arm $2'$ after its
    $s$-th pull in Stage III and $\theta_{2's} \leftarrow 0$, for $s = 0$;
14: while $t_2 \leq \log^2 T$ do
15:     if $t_2 \in \mathcal{T}_2'$ then
16:         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
17:             break
18:         end if
19:     end if
20:     $A_{t+1} = 2'$, $t \leftarrow t + 1$, $t_2 \leftarrow t_2 + 1$;
21: end while
22: ⋮
23: Stage IV
```

Theorem 5 *In the batched bandit problem, the expected number of rounds used in Algorithm 4 is $O(1)$. Moreover, the regret of Algorithm 4 is also asymptotically optimal, i.e., $\lim_{T\to\infty} R_\mu(T)/\log T = 2/\Delta$.*

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. It still remains an interesting open problem to achieve the minimax/instance-dependent optimality along with the asymptotic optimality and attain the optimal round cost at the same time.

Appendix B. Double Explore-then-Commit Algorithm for K-Armed Bandits

In this section, we extend our DETC framework to K-armed bandit problems. Due to the similarity in both structures and analyses between Algorithm 1 for known gaps and Algorithm 3 for unknown gaps, we only present the K-armed bandit algorithm for unknown gaps, which is more challenging.

Without loss of generality, we assume the best arm is 1. Let $\mu_1, \dots, \mu_K$ be the underlying mean reward of all K arms respectively. We denote $\Delta_i = \mu_1 - \mu_i$ as the gap between arm 1 and arm i for all $i \geq 2$. We assume $\Delta_i > 0$ for all $i \geq 2$. We present the double explore-then-commit algorithm for K-armed bandits in Algorithm 5. Similar to Algorithm 3 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 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' ($i \geq 2$) 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 f_{fail} initialized as 0, which will be set to 1 if any unchosen arm i' is played for $\log^2 T$ times; (4) in *Stage IV*, if $f_{\text{fail}} = 0$ and $\widehat{\mu}_{1'}$ is larger than the recalculated average reward for any other arm, then we play $1'$ till the end. Otherwise, $1'$ may not be the best arm. Then we play all arms $\log^2 T$ times, and play the arm with the largest recalculated average reward till the end.

Similar to the regret defined in (1), we define the regret for a bandit instance $\{\mu_1, \dots, \mu_K\}$ as $R_\mu(T) = T\max\{\mu_1, \mu_2\} - \mathbb{E}_\mu[\sum_{t=1}^T r_t]$. Now we present the regret bound of Algorithm 5.

Theorem 6 *The regret of Algorithm 5 with 1-subGaussian reward satisfies*

$$\lim_{T\to\infty} \frac{R_\mu(T)}{\log(T)} = \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}. \tag{9}$$

Remark 7 *In the second case of* Stage IV *of Algorithm 5, 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 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 6 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 5 is rather small and thus the extra ETC step does not affect the asymptotic regret of our DETC algorithm.*

Algorithm 5 Double Explore-then-Commit for K-Armed Bandits (DETC-K)

```
input $T$, $K$.

Stage I: Explore all arms uniformly
for $i = 1, 2 \cdots, K$ do
    Play every arm $\sqrt{\log T}$ times;
end for

Stage II: Commit the arm with the largest average reward
$1' \leftarrow \arg\max_i \widehat{\mu}_i$;
Play arm $1'$ $\log^2 T$ times and let $\mu'$ be the recalculated average reward for this play;

Stage III: Explore the unchosen arm in Stage II
$\{2', \cdots, K'\} = \{1, 2, \cdots, K\} \setminus \{1'\}$ and fail $\leftarrow 0$;
for $i = 2, 3, \cdots, K$ do
    $t_i \leftarrow 1$, $\theta_{i's}$ is the recalculated average reward of arm $i'$ after its $s$-th pull in Stage III and $\theta_{i'0} = 0$;
    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
        Play arm $i'$, $t_i \leftarrow t_i + 1$;
    end while
    if $t_i > \log^2 T$ then
        fail $\leftarrow 1$ and break;
    end if
end for

Stage IV: Commit the arm with the largest average reward after double exploration
$j' := \max_{i'} \theta_{i't_i}$;
if $\widehat{\mu}_{1'} \ge \theta_{j't_j}$ and fail $= 0$ then
    Let $a \leftarrow 1'$ and play arm $a$ till $T$ time step;
else
    Play every arm $\log^2 T$ times and let $a$ be the arm with the largest average reward for this play;
    Play arm $a$ till $T$ time steps.
end if
```

Appendix C. Proof of the Regret Bound of Algorithm 1

Now we are going to prove Theorem 1. We first present a technical lemma that characterizes the concentration properties of subGaussian random variables.

Lemma 8 (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 $\epsilon > 0$*

$$\mathbb{P}(\widehat{\mu} \ge \mu + \epsilon) \le \exp\bigg(-\frac{n\epsilon^2}{2\sigma^2}\bigg) \quad \textit{and} \quad \mathbb{P}(\widehat{\mu} \le \mu - \epsilon) \le \exp\bigg(-\frac{n\epsilon^2}{2\sigma^2}\bigg), \tag{10}$$

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

Proof [Proof of Theorem 1] Let τ_2 be the total number of times arm $2'$ is played 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 played in *Stage I*. Let $N_2(T)$ denote the total number of times

Algorithm 1 plays arm 2, which is calculated as

$$
\begin{aligned}
N_2(T) = \tau_1 + (T_1 - \tau_1)\,\mathbb{1}\{\widehat{\mu}_1(\tau_1) < \widehat{\mu}_2(\tau_1)\} + \tau_2\,\mathbb{1}\{\widehat{\mu}_1(\tau_1) \geq \widehat{\mu}_2(\tau_1)\} \\
+ (T - T_1 - \tau_1 - \tau_2)\,\mathbb{1}\{a = 2\}.
\end{aligned} \tag{11}
$$

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

$$
\begin{aligned}
R_\mu(T) &\leq \mathbb{E}\big[\Delta\tau_1 + \Delta(T_1 - \tau_1)\,\mathbb{1}\{\widehat{\mu}_1(\tau_1) < \widehat{\mu}_2(\tau_1)\} + \Delta\tau_2\,\mathbb{1}\{\widehat{\mu}_1(\tau_1) \geq \widehat{\mu}_2(\tau_1)\,\mathbb{1}\} + \Delta T\,\mathbb{1}\{a = 2\}\big] \\
&\leq \mathbb{E}\big[\Delta\tau_1 + \Delta T_1 \mathbb{P}(\widehat{\mu}_1(\tau_1) < \widehat{\mu}_2(\tau_1)) + \Delta\tau_2\mathbb{P}(\widehat{\mu}_1(\tau_1) \geq \widehat{\mu}_2(\tau_1)) + \Delta T\mathbb{P}(a = 2)\big] \\
&\leq \Delta\tau_1 + \underbrace{\Delta T_1 \mathbb{P}(\tau_1 < T_1, 1' = 2)}_{I_1} + \underbrace{\Delta\mathbb{E}[\tau_2]}_{I_2} + \underbrace{\Delta T\mathbb{P}(\tau_2 < T, a = 2)}_{I_3}.
\end{aligned} \tag{12}
$$

In what follows, we will bound these terms separately.

Bounding term I_1: Let X_i and Y_i be the rewards from playing 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 8 with any $\epsilon > 0$, we get

$$
\mathbb{P}(S_{\tau_1}/\tau_1 \leq \Delta - \epsilon) \leq \exp(-\tau_1\epsilon^2/4) \leq \exp(-\epsilon^2 \log(T_1\Delta^2)/\Delta^2), \tag{13}
$$

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

$$
I_1 = T_1\Delta\mathbb{P}(\tau_1 < T_1, 1' = 2) \leq 1/\Delta. \tag{14}
$$

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 8 and the union bound, $\mathbb{P}(E) \geq 1 - 2/(T\Delta^2)$. Therefore,

$$
\begin{aligned}
I_2 &= \Delta\mathbb{E}[\tau_2\,\mathbb{1}(E)] + \Delta\mathbb{E}[\tau_2\,\mathbb{1}(E^c)] \\
&= \Delta\mathbb{E}[\tau_2\,\mathbb{1}(E)] + \Delta\mathbb{E}[\tau_2 \mid E^c] \cdot \mathbb{P}(E^c) \\
&\leq \Delta\mathbb{E}[\tau_2\,\mathbb{1}(E)] + \Delta T \cdot \frac{2}{T\Delta^2} \\
&= \Delta\mathbb{E}[\tau_2\,\mathbb{1}(E, 1' = 1)] + \Delta\mathbb{E}[\tau_2\,\mathbb{1}(E, 1' = 2)] + 2/\Delta.
\end{aligned} \tag{15}
$$

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 played in *Stage III* of Algorithm 1. For ease of presentation, we define the following notations:

$$
Z_0 = 0, \quad Z_i = \mu' - Y_{i+\tau_1}, \quad S'_0 = 0, \quad S'_n = Z_1 + \cdots + Z_n, \tag{16}
$$

where $Y_{i+\tau_1}$ is the reward from playing 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 (Line 10), 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,

$$\begin{aligned}\left\{\tau_2 - 1 \geq \left\lceil \frac{\log(T\Delta^2) + x}{2(1-\epsilon_T)^2\Delta^2} \right\rceil \right\} &= \{\tau_2 - 1 \geq \lceil n_x \rceil\} \\ &\subseteq \left\{ S'_{\lceil n_x \rceil} \leq \frac{\log(T\Delta^2)}{2(1-\epsilon_T)\Delta} \right\}. \end{aligned} \tag{17}$$

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

$$\Delta' = \mu' - \mathbb{E}[Y_{1+\tau_1}] = \mu' - \mu_2 \geq \mu_1 - \epsilon_T\Delta - \mu_2 = (1-\epsilon_T)\Delta. \tag{18}$$

By Lemma 8, for any $\epsilon > 0$, we have

$$\mathbb{P}\left(\frac{S'_{\lceil n_x \rceil}}{\lceil n_x \rceil} \leq \Delta' - \epsilon \,\middle|\, E_1 \right) \leq \exp\left(-\lceil n_x \rceil \epsilon^2/2\right). \tag{19}$$

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 (19) yields

$$\begin{aligned}\mathbb{P}\left(S'_{\lceil n_x \rceil} \leq \frac{\log(T\Delta^2)}{2(1-\epsilon_T)\Delta} \,\middle|\, E_1 \right) &\leq \mathbb{P}\left(S'_{\lceil n_x \rceil} \leq \lceil n_x \rceil(\Delta' - \epsilon) \,\middle|\, E_1 \right) \\ &\leq \exp\left(- \frac{x^2}{4(\log(T\Delta^2) + x)} \right). \end{aligned} \tag{20}$$

This, when combined with (17), 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,

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

$$\begin{aligned}
&\le \frac{1}{2(1-\epsilon_T)^2\Delta^2}\int_{x_0}^{\infty}\exp\bigg(-\frac{x}{2\sqrt{\log(T\Delta^2)}+4}\bigg)\mathrm{d}x\\
&\le \frac{1}{2(1-\epsilon_T)^2\Delta^2}\int_{0}^{\infty}\exp\bigg(-\frac{x}{2\sqrt{\log(T\Delta^2)}+4}\bigg)\mathrm{d}x\\
&= \frac{\sqrt{\log(T\Delta^2)}+2}{(1-\epsilon_T)^2\Delta^2}. \qquad (21)
\end{aligned}$$

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

$$\begin{aligned}
\Delta\mathbb{E}[\tau_2 \mid E_1] &= \Delta\int_0^{\infty}\mathbb{P}(\tau_2 > v \mid E_1)\mathrm{d}v\\
&= \Delta\int_0^{n_{x_0}+2}\mathbb{P}(\tau_2 > v \mid E_1)\mathrm{d}v + \Delta\int_{n_{x_0}}^{\infty}\mathbb{P}(\tau_2 - 2 \ge v \mid E_1)\mathrm{d}v\\
&\le 2\Delta + \frac{\log(T\Delta^2)}{2(1-\epsilon_T)^2\Delta} + \frac{2\sqrt{\log(T\Delta^2)}+2}{(1-\epsilon_T)^2\Delta}. \qquad (22)
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\Delta\mathbb{E}[\tau_2\,\mathbb{1}(E, 1' = 1)] &= \Delta\mathbb{E}[\tau_2 \mid E_1]\cdot\mathbb{P}(E_1)\\
&\le \mathbb{P}(E_1)\cdot\bigg(2\Delta + \frac{\log(T\Delta^2)}{2(1-\epsilon_T)^2\Delta} + \frac{2\sqrt{\log(T\Delta^2)}+2}{(1-\epsilon_T)^2\Delta}\bigg). \qquad (23)
\end{aligned}$$

Let E_2 denote the event $\{E, 1' = 2\}$. In a manner similar to the proof of (22), we can show that

$$\begin{aligned}
\Delta\mathbb{E}[\tau_2\,\mathbb{1}(E, 1' = 2)] &= \Delta\mathbb{E}[\tau_2 \mid E_2]\cdot\mathbb{P}(E_2)\\
&\le \mathbb{P}(E_2)\cdot\bigg(2\Delta + \frac{\log(T\Delta^2)}{2(1-\epsilon_T)^2\Delta} + \frac{2\sqrt{\log(T\Delta^2)}+2}{(1-\epsilon_T)^2\Delta}\bigg). \qquad (24)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
I_2 &\le \Delta\mathbb{E}[\tau_2\,\mathbb{1}(E, 1' = 1)] + \Delta\mathbb{E}[\tau_2\,\mathbb{1}(E, 1' = 2)] + \frac{2}{\Delta}\\
&\le 2\Delta + \frac{2}{\Delta} + \frac{\log(T\Delta^2)}{2(1-\epsilon_T)^2\Delta} + \frac{2\sqrt{\log(T\Delta^2)}+2}{(1-\epsilon_T)^2\Delta}. \qquad (25)
\end{aligned}$$

Bounding term I_3: For term I_3, similar to (15), we have

$$\begin{aligned}
I_3 =& \Delta\cdot T\mathbb{P}[\tau_2 < T, a = 2 \mid E_1]\cdot\mathbb{P}[E_1]\\
&+ \Delta\cdot T\mathbb{P}[\tau_2 < T, a = 2 \mid E_2]\cdot\mathbb{P}[E_2] + \frac{2}{\Delta}. \qquad (26)
\end{aligned}$$

We will first prove that $\mathbb{P}(\tau_2 < T, a = 2 \mid E_1) \le 1/(T\Delta^2)$. 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 $\Delta' \ge (1-\epsilon_T)\Delta$ whenever E_1 occurs. Then,

$$\begin{aligned}
\mathbb{E}[\exp(-2\Delta(1-\epsilon_T)Z_1) \mid E_1] &= \mathbb{E}[\exp(-2\Delta(1-\epsilon_T)Z_1 + 2\Delta\Delta'(1-\epsilon_T) - 2\Delta\Delta'(1-\epsilon_T)) \mid E_1]\\
&= \mathbb{E}[\exp(-2\Delta(1-\epsilon_T)(Z_1 - \Delta') - 2\Delta\Delta'(1-\epsilon_T)) \mid E_1]
\end{aligned}$$

$$\begin{aligned}&\le \exp((-2(1-\epsilon_T)\Delta)^2/2 - 2(1-\epsilon_T)\Delta\Delta'))\\&\le \exp(2(1-\epsilon_T)\Delta((1-\epsilon_T)\Delta - \Delta'))\\&\le 1, \qquad (27)\end{aligned}$$

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 $n \ge 1$. 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 $\tau' = T \wedge \inf\{n > 1 : S'_n \le -\log(T\Delta^2)/(2\Delta(1-\epsilon_T))\}$ be a stopping time. Observe that conditioned on E_1,

$$\begin{aligned}\{\tau_2 < T, a = 2\} &\subseteq \left\{\exists 1 < n < T : S'_n \le -\frac{\log(T\Delta^2)}{2\Delta(1-\epsilon_T)}\right\}\\&= \{\tau' < T\}. \qquad (28)\end{aligned}$$

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

$$\begin{aligned}M_{\tau'} &= \exp(-2\Delta(1-\epsilon_T)S'_{\tau'})\\&\ge \exp(\log(T\Delta^2)) = T\Delta^2. \qquad (29)\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}(\tau_2 < T, a = 2 \mid E_1) &\le \mathbb{P}(\tau' < T \mid E_1)\\&\le \mathbb{P}(M_{\tau'} \ge T\Delta^2 \mid E_1)\\&\le \mathbb{E}[M_{\tau'}]/(T\Delta^2) \le 1/(T\Delta^2). \qquad (30)\end{aligned}$$

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

Completing the proof: Substituting (14), (25) and $I_3 \le 3/\Delta$ into (12) yields a total regret as follows

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

Recall the choice of ϵ_T in Theorem 1. By our choice that $T_1 = \lceil 2\log(T\Delta^2)/(\epsilon_T^2\Delta^2))\rceil$, we have

$$T_1 \le 1 + \max\{2\log^2 T, 8\log(T\Delta^2)/\Delta^2\}, \qquad (31)$$

which immediately implies, $\lim_{T\to\infty} 4\log(T_1\Delta^2)/(\Delta\log T) = 0$. Also note that $\lim_{T\to\infty}\epsilon_T = 0$. Thus, we have $\lim_{T\to\infty} R_\mu(T)/\log T = 1/(2\Delta)$. ■

Appendix D. Proof of the Regret Bound of Algorithm 3

Next, we provide the proof for Theorem 4. Note that the stopping time of *Stage I* in Algorithm 3 depends on the samples, and hence, the Hoeffding's inequality in Lemma 8 is not directly applicable. To address this issue, we provide the following variant of the *maximal inequality* (Feller, 2008).

Lemma 9 *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*

$$\mathbb{P}(\exists N \le n \le M, \widehat{\mu}_n + \gamma \le 0) \le \exp\bigg(-\frac{N\gamma^2}{2}\bigg). \tag{32}$$

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 10 *Let $\delta > 0$ be a constant 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 $T_1 \le T$,*

$$\sum_{n=1}^{T} \mathbb{P}\Bigg(\widehat{\mu}_n + \sqrt{\frac{4}{n}\log^+\bigg(\frac{T_1}{n}\bigg)} \ge \delta\Bigg) \le 1 + \frac{4\log^+(T_1\delta^2)}{\delta^2} + \frac{2}{\delta^2} + \frac{\sqrt{8\pi\log^+(T_1\delta^2)}}{\delta^2}; \tag{33}$$

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

$$\sum_{n=1}^{T} \mathbb{P}\Bigg(\widehat{\mu}_n + \sqrt{\frac{2}{n}\log\bigg(\frac{T}{n}\bigg(\log^2\frac{T}{n}+1\bigg)\bigg)} \ge \delta\Bigg)$$
$$\le 1 + \frac{2\log(T\delta^2(\log^2(T\delta^2)+1))}{\delta^2} + \frac{2}{\delta^2} + \frac{\sqrt{4\pi\log(T\delta^2(\log^2(T\delta^2)+1))}}{\delta^2}; \tag{34}$$

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

$$\mathbb{P}\Bigg(\exists s \le T : \widehat{\mu}_s + \sqrt{\frac{2}{s}\log\bigg(\frac{T}{s}\bigg(\log^2\frac{T}{s}+1\bigg)\bigg)} + \delta \le 0\Bigg) \le \frac{4(16e^2+1)}{T\delta^2}. \tag{35}$$

Proof [Proof of Theorem 4] Let τ_1 be the number of times each arm is played in *Stage I* of Algorithm 3 and τ_2 be the total number of times arm $2'$ is played in *Stage III* of Algorithm 3. Similar to (12), the regret of Algorithm 3 can be decomposed as follows

$$R_\mu(T) \le \underbrace{\Delta T_1\mathbb{P}(\tau_1 < T_1, 1' = 2)}_{I_1} + \underbrace{\Delta\mathbb{E}[\tau_1] + \Delta\mathbb{E}[\tau_2]}_{I_2} + \underbrace{\Delta T\mathbb{P}(\tau_2 < T, a = 2)}_{I_3}. \tag{36}$$

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, $s = 1, 2, \ldots$. Let $Z_s = (X_s - Y_s - \Delta)/\sqrt{2}$. Then Z_s is a 1-subGaussian random variable with zero mean. Let $S_s = \sum_{i=1}^{s} Z_i$. Recall that $\widehat{\mu}_{k,s}$ is the average reward for arm k after its s-th pull. Applying the standard peeling technique, we have

$$\begin{aligned}
\mathbb{P}(\tau_1 < T_1, 1' = 2) &\le \mathbb{P}\bigg(\exists s \in \mathbb{N} : 2s \le T,\ \widehat{\mu}_{1,s} - \widehat{\mu}_{2,s} \le -\sqrt{\frac{8\log^+(T_1/(2s))}{s}}\bigg)\\
&\le \mathbb{P}\bigg(\exists s \ge 1 : \frac{\sum_{i=1}^{s} Z_i}{s} \le -\sqrt{\frac{4\log^+(T_1/(2s))}{s}} - \frac{\Delta}{\sqrt{2}}\bigg)\\
&\le \sum_{j=0}^{\infty} \mathbb{P}\bigg(\exists s \in [2^j, 2^{j+1}] : \frac{\sum_{i=1}^{s} Z_i}{s} + \sqrt{\frac{4\log^+(T_1/(2s))}{s}} + \frac{\Delta}{\sqrt{2}} \le 0\bigg)\\
&\le \sum_{j=0}^{\infty} \mathbb{P}\bigg(\exists s \in [2^j, 2^{j+1}] : \frac{\sum_{i=1}^{s} Z_i}{s} + \sqrt{\frac{4\log^+(T_1/2^{j+2})}{2^{j+1}}} + \frac{\Delta}{\sqrt{2}} \le 0\bigg)\\
&\le \sum_{j=0}^{\infty} \exp\bigg(-2^{j-1}\bigg(\sqrt{\frac{\log^+(T_1/2^{j+2})}{2^{j-1}}} + \frac{\Delta}{\sqrt{2}}\bigg)^2\bigg),
\end{aligned}$$

where the last inequality comes from Lemma 9. Therefore, we have

$$\begin{aligned}
\mathbb{P}(\tau_1 < T_1, 1' = 2) &\le \sum_{j=0}^{\infty} \exp\bigg(-\log^+\bigg(\frac{T_1}{2^{j+2}}\bigg) - 2^{j-2}\Delta^2\bigg)\\
&= \frac{1}{T_1}\sum_{j=0}^{\infty} 2^{j+2}\exp(-2^{j-2}\Delta^2)\\
&\le \frac{16}{eT_1\Delta^2} + \frac{1}{T_1}\int_0^{\infty} 2^{j+2}\exp(-2^{j-2}\Delta^2)\mathrm{d}j\\
&\le \frac{16}{eT_1\Delta^2} + \frac{-16}{\log 2\cdot T_1\Delta^2}\exp(-2^{j-2}\Delta^2)\bigg|_0^{\infty}\\
&\le \frac{30}{T_1\Delta^2},
\end{aligned} \tag{37}$$

where the first inequality we used the factor that $(x+y)^2 \ge x^2 + y^2$, $x, y \ge 0$, the third inequality follows the fact that the integral function has a maximum value $16/(eT_1\Delta^2)$ and for such function we have $\sum_{j=0}^{\infty} f(j) \le \max_{j\in[0,\infty)} f(j) + \int_0^{\infty} f(j)\mathrm{d}j$. Thus, we have proved that $\Delta T_1 \mathbb{P}(\tau_1 < T_1, 1' = 2) \le 30/\Delta$.

Bounding term I_2: By the definition of τ_1 and the stopping rule of *Stage I* in Algorithm 3, we have

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

$$\leq 1 + \frac{8\log^{+}(T_1\Delta^2/4)}{\Delta^2} + \frac{4}{\Delta^2} + \frac{2\sqrt{8\pi\log^{+}(T_1\Delta^2/4)}}{\Delta^2}, \tag{38}$$

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 10 since $-Z_i$ are 1-subGaussian variables as well.

Recall that $\epsilon_T = \sqrt{2\log(T\Delta^2)/(T_1\Delta^2)}$ and $\epsilon_T \in (0, 1/2)$. Let E be the event $\mu' \in [\mu_{1'} - \epsilon_T\Delta, \mu_{1'} + \epsilon_T\Delta]$. Applying Lemma 8 and union bound, $\mathbb{P}(E) \geq 1 - 2/(T\Delta^2)$. Similar to (15), we have

$$\mathbb{E}[\tau_2] \leq \mathbb{E}[\tau_2\,\mathbb{1}(E, 1' = 1)] + \mathbb{E}[\tau_2\,\mathbb{1}(E, 1' = 2)] + 2/\Delta^2. \tag{39}$$

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 $T\Delta^2 \geq 16e^3$, we have that conditioned on E_1, $T(\Delta')^2 \geq (1-\epsilon_T)^2 T\Delta^2 \geq 4e^3$. 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 3, it holds that

$$\begin{aligned}
\mathbb{E}[\tau_2 \mid E_1] &\leq \sum_{t_2=1}^{T} \mathbb{P}(\tau_2 \geq t_2 \mid E_1) \\
&= \sum_{t_2=1}^{T} \mathbb{P}\bigg(\mu' - \theta_{2',t_2} \leq \sqrt{\frac{2}{t_2}\log\Big(\frac{T}{t_2}\Big(\log^2\frac{T}{t_2} + 1\Big)\Big)}\ \bigg|\ E_1\bigg) \\
&= \sum_{t_2=1}^{T} \mathbb{P}\bigg(-\frac{\sum_{i=1}^{t_2} W_i}{t_2} + \sqrt{\frac{2}{t_2}\log\Big(\frac{T}{t_2}\Big(\log^2\frac{T}{t_2} + 1\Big)\Big)} \geq \Delta'\ \bigg|\ E_1\bigg) \\
&\leq 1 + \frac{2 + 2\log(4T\Delta^2(\log^2(4T\Delta^2)+1)) + \sqrt{4\pi\log(4T\Delta^2(\log^2(4T\Delta^2)+1))}}{(1-\epsilon_T)^2\Delta^2}.
\end{aligned} \tag{40}$$

where the last inequality is due to the second statement of Lemma 10 and $-W_i$ are 1-subGuassian. Let $E_2 = \{E, 1' = 2\}$, using the same argument, we can derive same bound as in (40) for $\mathbb{E}[\tau_2 \mid E_2]$. Then We have

$$\begin{aligned}
\Delta\mathbb{E}[\tau_2] &\leq \Delta\mathbb{E}[\tau_2\,\mathbb{1}(E_1)] + \Delta\mathbb{E}[\tau_2\,\mathbb{1}(E_2)] + \frac{2}{\Delta} \\
&= \Delta\mathbb{E}[\tau_2|E_1]\mathbb{P}(E_1) + \Delta\mathbb{E}[\tau_2|E_2]\mathbb{P}(E_2) + \frac{2}{\Delta} \\
&\leq \Delta + \frac{2}{\Delta} + \frac{2 + 2\log(4T\Delta^2(\log^2(4T\Delta^2)+1)) + \sqrt{4\pi\log(4T\Delta^2(\log^2(4T\Delta^2)+1))}}{(1-\epsilon_T)^2\Delta}.
\end{aligned} \tag{41}$$

Bounding term I_3: Similar to (39),

$$I_3 \leq \Delta T\mathbb{P}[\tau_2 < T, a = 2 \mid E_1]\mathbb{P}[E_1] + \Delta T\mathbb{P}[\tau_2 < T, a = 2 \mid E_2]\mathbb{P}[E_2] + \frac{2}{\Delta}. \tag{42}$$

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 10, we have

$$\begin{aligned}\mathbb{P}(\tau_2 < T, a = 2 \mid E_1) &\leq \mathbb{P}\bigg(\exists t_2 \geq 1, \mu' - \theta_{2',t_2} + \sqrt{\frac{2}{t_2}\log\Big(\frac{T}{t_2}\Big(\log^2\frac{T}{t_2}+1\Big)\Big)} \leq 0 \,\bigg|\, E_1\bigg)\\ &\leq \mathbb{P}\bigg(\exists t_2 \geq 1, \mu' - \theta_{2',t_2} - \Delta' + \Delta' + \sqrt{\frac{2}{t_2}\log\Big(\frac{T}{t_2}\Big(\log^2\frac{T}{t_2}+1\Big)\Big)} \leq 0 \,\bigg|\, E_1\bigg)\\ &\leq \frac{4(16e^2+1)}{T(1-\epsilon_T)^2\Delta^2}. \qquad (43)\end{aligned}$$

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 (43) for term $\mathbb{P}(\tau_2 < T, a = 2 \mid E_2)$.
Therefore,

$$I_3 \leq \frac{2}{\Delta} + \frac{4(16e^2+1)}{(1-\epsilon_T)^2\Delta}. \qquad (44)$$

Completing the proof: Therefore, substituting (37), (38), (41) and (44) into (36), we have

$$\begin{aligned}R_\mu(T) \leq 2\Delta &+ \frac{38 + 8\log^+(T_1\Delta^2/4) + 2\sqrt{8\pi\log^+(T_1\Delta^2)}}{\Delta}\\ &+ \frac{4(16e^2+2) + 2\log(4T\Delta^2(\log^2(4T\Delta^2)+1)) + \sqrt{4\pi\log(4T\Delta^2(\log^2(4T\Delta^2)+1))}}{(1-\epsilon_T)^2\Delta}.\end{aligned}$$

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

Appendix E. Proof of the 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 9

Our proof relies on the following maximal inequality for supermartingales.

Lemma 11 (Ville (1939)) *If (S_n) is a non-negative supermartingale, then for any $x > 0$,*

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

Proof [Proof of Lemma 9] 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 (32) holds trivially. Otherwise, if event $\{\exists N \leq n \leq M, \widehat{\mu}_n + \gamma \leq 0\}$ holds, then the following three inequalities also hold simultaneously:

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

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

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

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

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

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

E.2. Proof of Lemma 10

To prove Lemma 10, we also need the following technical lemma from Ménard and Garivier (2017).

Lemma 12 *For all $\beta > 1$ we have*

$$\frac{1}{e^{\log(\beta)/\beta} - 1} \leq 2\max\{\beta, \beta/(\beta-1)\}. \tag{45}$$

Proof [Proof of Lemma 10] For the first statement, let $\gamma' = 4\log^+(T_1\delta^2)/\delta^2$. Note that for $n \geq \gamma'$, it holds that $n\delta^2 \geq 4$ and

$$\delta\sqrt{\frac{\gamma'}{n}} = \sqrt{\frac{4}{n}\log^+(T_1\delta^2)} \geq \sqrt{\frac{4}{n}\log^+\bigg(\frac{T_1}{n}\bigg)}. \tag{46}$$

Therefore, using the same argument in (49) we can show that where we used the fact that $T\delta^2 \geq e^2$ and hence $\delta^2 = 2\log(T\delta^2(\log^2(T\delta^2)+1))/\gamma \geq 1/\gamma \geq 1/n$. Therefore, we have

$$\begin{aligned}
\sum_{n=1}^{T}\mathbb{P}\bigg(\widehat{\mu}_n + \sqrt{\frac{4}{n}\log^+\bigg(\frac{T_1}{n}\bigg)} \geq \delta\bigg) &\leq \gamma' + \sum_{n=\lceil\gamma\rceil}^{T}\mathbb{P}\bigg(\widehat{\mu}_n \geq \delta\bigg(1 - \sqrt{\frac{\gamma'}{n}}\bigg)\bigg) \\
&\leq \gamma + \sum_{n=\lceil\gamma\rceil}^{\infty}\exp\bigg(-\frac{n\delta^2}{2}\bigg(1 - \sqrt{\frac{\gamma}{n}}\bigg)^2\bigg)
\end{aligned} \tag{47}$$

$$
\begin{aligned}
&\leq \gamma' + \sum_{n=\lceil \gamma' \rceil}^{\infty} \exp\bigg(-\frac{\delta^2(\sqrt{n}-\sqrt{\gamma'})^2}{2}\bigg) && (48)\\
&\leq \gamma' + 1 + \int_{\gamma'}^{\infty} \exp\bigg(-\frac{\delta^2(\sqrt{x}-\sqrt{\gamma'})^2}{2}\bigg)\mathrm{d}x\\
&\leq \gamma' + 1 + \frac{2}{\delta}\int_0^{\infty}\Big(\frac{y}{\delta}+\sqrt{\gamma'}\Big)\exp(-y^2/2)\mathrm{d}y\\
&\leq \gamma' + 1 + \frac{2}{\delta^2} + \frac{\sqrt{2\pi\gamma'}}{\delta}, && (49)
\end{aligned}
$$

where (48) is the result of Lemma 8 and (49) is due to the fact that $\int_0^\infty y\exp(-y^2/2)\mathrm{d}y = 1$ and $\int_0^\infty \exp(-y^2/2)\mathrm{d}y = \sqrt{2\pi}/2$. (49) immediately implies the claim in the first statement:

$$
\begin{aligned}
\sum_{n=1}^{T}\mathbb{P}\bigg(\widehat{\mu}_n + \sqrt{\frac{4}{n}\log^+\Big(\frac{T_1}{n}\Big)} \geq \delta\bigg) &\leq \gamma' + \sum_{n=\lceil\gamma'\rceil}^{T}\mathbb{P}\bigg(\widehat{\mu}_n \geq \delta\bigg(1-\sqrt{\frac{\gamma'}{n}}\bigg)\bigg)\\
&\leq \gamma' + 1 + \frac{2}{\delta^2} + \frac{\sqrt{2\pi\gamma'}}{\delta}. && (50)
\end{aligned}
$$

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

$$
\sum_{n=1}^{T}\mathbb{P}\bigg(\widehat{\mu}_n + \sqrt{\frac{4}{n}\log^+\Big(\frac{T_1}{n}\Big)} \geq \delta\bigg) \leq 1 + \frac{4\log^+(T_1\delta^2)}{\delta^2} + \frac{2}{\delta^2} + \frac{\sqrt{8\pi\log^+(T_1\delta^2)}}{\delta^2}. \quad (51)
$$

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

$$
\gamma = \frac{2\log(T\delta^2(\log^2(T\delta^2)+1))}{\delta^2}. \quad (52)
$$

Note that for all $n \geq \gamma$, it holds that

$$
\delta\sqrt{\frac{\gamma}{n}} = \sqrt{\frac{2\log(T\delta^2(\log^2(T\delta^2)+1))}{n}} \geq \sqrt{\frac{2}{n}\log\bigg(\frac{T}{n}\bigg(\log^2\frac{T}{n}+1\bigg)\bigg)}, \quad (53)
$$

where we used the fact that $T\delta^2 \geq e^2$ and hence $\delta^2 = 2\log(T\delta^2(\log^2(T\delta^2)+1))/\gamma \geq 1/\gamma \geq 1/n$. Therefore, using the same argument in (49) we can show that

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

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 $\{\exists s : f(\delta) \leq s \leq T\}$.

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

$$\leq \underbrace{\mathbb{P}\bigg(\exists s \leq f(\delta) : \widehat{\mu}_s \leq -\sqrt{\frac{2}{s}\log\Big(\frac{T}{s}\Big(\log^2\frac{T}{s}+1\Big)\Big)}\bigg)}_{A_1} + \underbrace{\mathbb{P}(\exists s, f(\delta) \leq s \leq T : \widehat{\mu}_s \leq -\delta)}_{A_2}. \tag{54}$$

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.

$$A_1 \leq \sum_{\ell=0}^{\infty} \underbrace{\mathbb{P}\bigg(\exists s, \frac{f(\delta)}{\beta^{\ell+1}} \leq s \leq \frac{f(\delta)}{\beta^{\ell}} : \widehat{\mu}_s + \sqrt{\frac{2}{s}\log\Big(\frac{T}{s}\Big(\log^2\frac{T}{s}+1\Big)\Big)} \leq 0\bigg)}_{A_1^{\ell}}. \tag{55}$$

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

$$\gamma_\ell = \frac{\beta^\ell}{f(\delta)}\log\bigg(\frac{T\beta^\ell}{2f(\delta)}\Big(1+\log^2\frac{T}{2f(\delta)}\Big)\bigg), \tag{56}$$

which by definition immediately implies

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

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

$$\begin{aligned}
&\mathbb{P}\bigg(\exists s, \frac{f(\delta)}{\beta^{\ell+1}} \leq s \leq \frac{f(\delta)}{\beta^{\ell}} : \widehat{\mu}_s + \sqrt{\frac{2}{s}\log\Big(\frac{T}{s}\Big(\log^2\frac{T}{s}+1\Big)\Big)} \leq 0\bigg)\\
&\leq \mathbb{P}\bigg(\exists \frac{f(\delta)}{\beta^{\ell+1}} \leq s \leq \frac{f(\delta)}{\beta^{\ell}} : \widehat{\mu}_s + \sqrt{2\gamma_\ell} \leq 0\bigg)\\
&\leq \exp\bigg(-\frac{f(\delta)}{\beta^{\ell+1}}\gamma_\ell\bigg)\\
&= e^{-\ell\log(\beta)/\beta - C/\beta},
\end{aligned} \tag{57}$$

where the second inequality is by Doob's maximal inequality (Lemma 9), the last equation is due to the definition of γ_ℓ, and the parameter C is defined to be

$$C := \log\bigg(\frac{T}{2f(\delta)}\Big(1+\log^2\frac{T}{2f(\delta)}\Big)\bigg). \tag{58}$$

Substituting (57) back into (55), 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 12. 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

$$C = \log\bigg(\frac{T}{2f(\delta)}\Big(1+\log^2\frac{T}{2f(\delta)}\Big)\bigg) \geq \log\bigg(\frac{T}{2f(\delta)}\bigg) = \log\bigg(\frac{T\delta^2}{4\log(\frac{T\delta^2}{4})}\bigg) \geq 3/2. \tag{59}$$

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 $A_1 \leq 4eCe^{-C/\beta} = 4e^2Ce^{-C}$. Note that

$$\begin{aligned} Ce^{-C} &= \bigg(\frac{T}{2f(\delta)}\bigg(1+\log^2\frac{T}{2f(\delta)}\bigg)\bigg)^{-1}\log\bigg(\frac{T}{2f(\delta)}\bigg(1+\log^2\frac{T}{2f(\delta)}\bigg)\bigg) \\ &\leq \frac{2f(\delta)}{T\log^2(T/(2f(\delta)))}\log\bigg(\frac{T}{2f(\delta)}\bigg(1+\log^2\frac{T}{2f(\delta)}\bigg)\bigg) \\ &\leq \frac{4f(\delta)}{T\log(T/(2f(\delta)))} \\ &= \frac{8\log(T\delta^2/4)}{T\delta^2\log([T\delta^2/4]/\log(T\delta^2/4))} \\ &\leq \frac{16}{T\delta^2}, \end{aligned} \tag{60}$$

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

$$\frac{\log(x(1+\log^2 x))}{\log x} \leq 2 \qquad \text{and} \qquad \frac{\log x}{\log(x/\log x)} \leq 2. \tag{61}$$

Therefore, we have proved so far $A_1 \leq 64e^2/(T\delta^2)$. For term A_2 in (54), we can again apply the maximal inequality in Lemma 9 and obtain

$$A_2 = \mathbb{P}(\exists s, f(\delta) \leq s \leq T : \widehat{\mu}_s \leq -\delta) \leq e^{-\delta^2 f(\delta)/2} = \frac{4}{T\delta^2}. \tag{62}$$

Finally, combining the above results, we get

$$\mathbb{P}\Bigg(\exists s \leq f(\delta), \widehat{\mu}_s + \sqrt{\frac{2}{s}\log\bigg(\frac{T}{s}\bigg(\log^2\frac{T}{s}+1\bigg)\bigg)} + \delta \leq 0\Bigg) \leq \frac{4(16e^2+1)}{T\delta^2}. \tag{63}$$

This completes the proof. ■

Appendix F. Round Complexity of DETC for Batched Bandit Models

In this section, we derive the round complexities of Algorithms 1 and 3 for batched bandit models. We will prove that 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.

F.1. Proof of Theorem 2

We first prove the round complexity for DETC (Algorithm 1) when the gap Δ is known.

Proof The analysis is very similar to that of Theorem 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 (20), we have

$$\begin{aligned} \mathbb{P}(\tau_2 > n_{x_i} \mid E_1) \le \mathbb{P}\bigg(S_{n_{x_i}} \le \frac{\log(T\Delta^2)}{2(1-\epsilon_T)\Delta} \,\bigg|\, E_1\bigg) &\le \exp\bigg(-\frac{x_i^2}{4(\log(T\Delta^2) + x_i)}\bigg) \\ &\le \exp\bigg(-\frac{x_i}{2\sqrt{\log(T\Delta^2)} + 4}\bigg) \\ &\le 2^{-i}. \end{aligned} \tag{64}$$

Thus, the expected number of rounds of queries needed in *Stage III* of Algorithm 1 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 $\mathbb{P}(E^c) \le 2/(T\Delta^2)$. 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 3 is $O(1)$.

Following the same proof in (21) and (22), 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 (22). The bounds for other terms remain the same. Therefore, the batched version of Algorithm 1 is still asymptotically optimal, instance-dependent optimal and minimax optimal. ■

F.2. Proof of Theorem 5

Now we prove the round complexity and regret bound for DETC (Algorithm 3) when the gap Δ is unknown.

Proof For the sake of simplicity, we use the same notations that are used in Theorem 4 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

$$\sqrt{\log T} \ge 16 \log^+(T_1\Delta^2/2)/\Delta^2, \tag{65}$$

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 (65), 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

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

$$
\begin{aligned}
&\le \exp(-i) \\
&\le 2^{-i},
\end{aligned}
$$

where the third inequality follows from Lemma 8 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 (7), the expected number of rounds needed in *Stage I* of Algorithm 3 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 (38).

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 8, 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

$$
\sqrt{\frac{4\log(T\Delta^2)}{\Delta^2\log^2 T}} \le \frac{1}{(\log T)^{\frac{1}{3}}}, \tag{66}
$$

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

$$
\sqrt{\log T}(\Delta')^2 \ge 2(\log\log T)^2. \tag{67}
$$

Applying Lemma 8, we have

$$
\begin{aligned}
\mathbb{P}\Big(\mu_{2'} - \Delta'(\log T)^{-\frac{1}{4}} \le \theta_{2',N_1} \le \mu_{2'} + \Delta'(\log T)^{-\frac{1}{4}} \mid E_1\Big) &\ge 1 - 2\exp\bigg(-\frac{2\log T(\Delta')^2}{2\sqrt{\log T}\log\log T}\bigg) \\
&\ge 1 - \frac{2}{\log^2 T},
\end{aligned} \tag{68}
$$

where the last inequality follows by (67). This means that after the first round of *Stage III* in Algorithm 3, 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

$$
\begin{aligned}
\mathbb{E}[\tau_2] &\le \mathbb{E}[\tau_2 \mid E_1, E_3]\mathbb{P}[E_1, E_3] + \mathbb{E}[\tau_2 \mid E_2, E_3]\mathbb{P}[E_2, E_3] + \mathbb{E}[\tau_2 \mid E_3^c]\mathbb{P}[E_3^c] + \mathbb{E}[\tau_2 \mid E^c]\mathbb{P}[E^c] \\
&\le \mathbb{E}[\tau_2 \mid H_1]\mathbb{P}[H_1] + \mathbb{E}[\tau_2 \mid H_2]\mathbb{P}[H_2] + \mathbb{E}[\tau_2 \mid E_3^c]\mathbb{P}[E_3^c] + 2/(T\Delta^3)
\end{aligned} \tag{69}
$$

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 (8), 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

$$\log^3 T \geq (\Delta')^2(\log^2(T(\Delta')^2) + 1). \tag{70}$$

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

$$\frac{2(1 + 1/\sqrt[4]{\log T})^2 \log(T \log^3 T)}{\widehat{\Delta}^2} \geq \frac{2\log(T\log^3 T)}{(\Delta')^2} \geq \gamma', \tag{71}$$

where the last inequality is due to (70). On the other hand, we also have that conditioned on H_1

$$s'_i \geq \frac{2(1 + 1/\sqrt[4]{\log T})^2 \log(T \log^3 T)}{\widehat{\Delta}^2} \geq \frac{2}{(\Delta')^2}. \tag{72}$$

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

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

Recall the definition $W_i = \mu' - Y_{i+\tau_1} - \Delta'$ used in (40). From the stopping rule of *Stage III* in Algorithm 3, conditioned on H_1, we obtain

$$\begin{aligned}
\mathbb{P}(\tau_2 \geq s'_i \mid H_1) &\leq \mathbb{P}\bigg(\mu' - \theta_{2',s'_i} \leq \sqrt{\frac{2}{s'_i}\log\bigg(\frac{T}{s'_i}\bigg(\log^2\frac{T}{s'_i} + 1\bigg)\bigg)}\ \bigg|\ H_1\bigg)\\
&= \mathbb{P}\bigg(\frac{\sum_{i=1}^{s'_i} W_i}{s'_i} + \Delta' \leq \sqrt{\frac{2}{s'_i}\log\bigg(\frac{T}{s'_i}\bigg(\log^2\frac{T}{s'_i} + 1\bigg)\bigg)}\ \bigg|\ H_1\bigg)\\
&\leq \exp\bigg(-\frac{s'_i(\Delta')^2}{2}\bigg(1 - \sqrt{\frac{\gamma'}{s'_i}}\bigg)^2\bigg)\\
&= \exp\bigg(-\frac{(\Delta')^2}{2}(\sqrt{s'_i} - \sqrt{\gamma'})^2\bigg)\\
&= \exp\bigg(-\frac{(\Delta')^2}{2}\bigg(\frac{s'_i - \gamma'}{\sqrt{s'_i} + \sqrt{\gamma'}}\bigg)^2\bigg)\\
&\leq \exp\bigg(-\frac{i^2(\log T)^{4/3}}{8s'_i(\Delta')^2}\bigg),
\end{aligned} \tag{73}$$

where the second inequality from Lemma 8 and in the last inequality we used the fact that $s'_i - \gamma' \geq i(1 + 1/\sqrt[4]{\log T})^2(\log T)^{\frac{2}{3}}/(\widehat{\Delta}^2) \geq i(\log T)^{\frac{2}{3}}/(\Delta')^2$ by (71). Choose sufficiently large T to ensure

$$(\log T)^{\frac{4}{3}} \geq 8s'_i(\Delta')^2. \tag{74}$$

Substituting (74) back into (73) yields $\mathbb{P}(\tau_2 \geq s'_i \mid H_1) \leq 1/2^i$. Similarly, $\mathbb{P}(\tau_2 \geq s'_i \mid H_2) \leq 1/2^i$, Thus conditioned on H_1 (or H_2), the expected rounds used in *Stage III* of Algorithm 3 is upper bounded by $\sum_{i=1}^{\infty} i/2^i \le 2$. Recall that from (66), $\epsilon'_T \le 1/(\log T)^{\frac{1}{3}}$. Conditional on H_1, the expectation of τ_2 is upper bounded by

$$\begin{aligned}
\mathbb{E}[\tau_2 \mid H_1] &\le s'_1 + \sum_{i=2}[(s'_i - s'_1)\mathbb{P}(\tau_2 \ge s'_i \mid H_1)] \\
&\le \frac{2(1+1/(\log T)^{\frac{1}{4}})^2 \log(T\log^3 T)}{\widehat{\Delta}^2} + \frac{2(1+1/\sqrt[4]{\log T})^2(\log T)^{\frac{2}{3}}}{\widehat{\Delta}^2} \\
&\le \frac{2(1+1/(\log T)^{\frac{1}{4}})^2 \log(T\log^3 T) + 2(1+1/(\log T)^{\frac{1}{4}})^2(\log T)^{\frac{2}{3}}}{(1-1/(\log T)^{\frac{1}{3}})^2(1-1/(\log T)^{\frac{1}{4}})^2\Delta^2},
\end{aligned} \tag{75}$$

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 (75) 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 (68). Therefore $\mathbb{E}[\tau_2 \mid E_3^c]$ can be upper bounded by 2, which is dominated by (75). 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 4 arises when we terminate *Stage III* with $t_2 = \log^2 T$. The term I_3 can be written as

$$I_3 = \Delta T \mathbb{P}(\tau_2 = \log^2 T, a=2) + \Delta T \mathbb{P}(\tau_2 < \log^2 T, a=2), \tag{76}$$

We can derive same bound as (44) 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 $n \ge 1$. Then $X_i - Y_{i+\tau_1} - \Delta$ is a $\sqrt{2}$-subGaussian random variable. Applying Lemma 8 with $\epsilon = \Delta$ yields

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

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

$$\exp(-\Delta^2 \log^2 T/4) \le \frac{1}{T}. \tag{77}$$

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

$$\mathbb{P}(a=2) \le \frac{1}{T}. \tag{78}$$

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

To summarize, we can choose a sufficiently large T such that all the conditions (65), (66), (67), (70), (74) and (77) are satisfied simultaneously. Then the round complexity of Algorithm 3 is $O(1)$. For the regret bound, since the only difference between ALgorithm 4 and Algorithm 3 is the stopping rules of *Stage I* and *Stage III*, we only need to combine the regret for terms (75) and (78) and the fact that $\Delta\mathbb{E}[\tau_1] \le 4\Delta\sqrt{\log T}$ to obtain the total regret. Therefore, we have $\lim_{T\to\infty} R(T)/\log T = 2/\Delta$. ■

Appendix G. Proof of the Asymptotically Optimal Regret Bound of DETC-K

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

Proof [Proof of Theorem 6] 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

$$\lim_{T\to\infty} \frac{\mathbb{E}[T_i]}{\log(T)} = \frac{2}{\Delta_i^2}. \tag{79}$$

Denote $\tau_{2,i}$ as the number of pulls of arm i in *Stage III* of Algorithm 5. Similar to (11) and (12), the term $\mathbb{E}[T_i]$ can be decomposed as follows

$$\begin{aligned}\mathbb{E}[T_i] \leq \sqrt{\log T} + \underbrace{\log^2 T\mathbb{P}(1' = i)}_{I_1} + \underbrace{\mathbb{E}[\tau_{2,i}]}_{I_2} + \underbrace{T\mathbb{P}(\widehat{\mu}_{1'} \geq \theta_{j',t_j}, \text{fail=0}, a = i)}_{I_3} \\ + \underbrace{\log^2 T\mathbb{P}(\text{fail=1}) + T\mathbb{P}(\text{fail=1}, a = i)}_{I_4},\end{aligned} \tag{80}$$

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 5.

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 playing arm 1 and arm i τ_1 times, using Lemma 8, we get

$$\mathbb{P}(S_{\tau_1}^i/\tau_1 \leq \Delta_i - \epsilon) \leq \exp(-\tau_1\epsilon^2/4). \tag{81}$$

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

$$\frac{\sqrt{\log T}}{\log K + 2\log\log T} \geq \frac{4}{\Delta_i^2}, \tag{82}$$

Setting $\epsilon = \Delta_i$ in (81), 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

$$\mathbb{P}(\widehat{\mu}_{1,\tau_1} \geq \max_i \widehat{\mu}_{i,\tau_1}) = \mathbb{P}(1' = 1) \geq 1 - \frac{1}{\log^2 T}, \tag{83}$$

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 8, we have

$$\mathbb{P}(\mu' \notin (\mu_{1'} - \epsilon_i\Delta_i, \mu_{1'} + \epsilon_i\Delta_i)) \leq 2/(T\Delta_i^2). \tag{84}$$

Similar to (66), we choose a large T such that for all $\Delta_i > 0$,

$$\sqrt{\frac{4\log(T\Delta_i^2)}{\Delta_i^2\log^2 T}} \leq \frac{1}{(\log T)^{\frac{1}{3}}}, \tag{85}$$

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

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

$$\begin{aligned}&\le \mathbb{E}[\tau_{2,i}\,\mathbb{1}(1'=1)]+1\\&\le \mathbb{E}[\tau_{2,i}\,\mathbb{1}(E_1)]+\mathbb{E}[\tau_{2,i}\,\mathbb{1}(E^c)]+1\\&\le 1+\frac{2}{\Delta_i}+\mathbb{E}[\tau_{2,i}\mid E_1].\end{aligned}\tag{86}$$

We can derive the same bound as $\mathbb{E}[\tau_2 \mid E_1]$ in (40) for $\mathbb{E}[\tau_{2,i} \mid E_1]$. We have

$$I_2=\mathbb{E}[\tau_{2,i}\mid E_1]\le 1+\frac{2+2\log(4T\Delta_i^2(\log^2(4T\Delta_i^2)+1))+\sqrt{4\pi\log(4T\Delta_i^2(\log^2(4T\Delta_i^2)+1))}}{(1-\epsilon_i)^2\Delta_i^2}.\tag{87}$$

Bounding term I_3: When fail $=0$, we can follow the same proof for bounding I_3 in (43). Therefore, we can obtain

$$I_3\le\frac{2}{\Delta_i^2}+\frac{4(16e^2+1)}{(1-\epsilon_i)^2\Delta_i^2}.\tag{88}$$

Bounding term I_4: For term $\mathbb{P}(\text{fail=1})$, similar to (86), we have

$$\begin{aligned}\mathbb{P}(\text{fail=1})&=\mathbb{P}(\text{fail=1}\mid 1'=1)\Pr(1'=1)+\mathbb{P}(\text{fail=1}\mid 1'\neq 1)\Pr(1'\neq 1)\\&\le \mathbb{P}(\text{fail=1}\mid 1'=1)+\frac{1}{\log^2 T}\\&\le \mathbb{P}(\text{fail=1}\mid E,1'=1)\Pr(E\mid 1'=1)+\Pr(E^c\mid 1'=1)+\frac{1}{\log^2 T}\\&\le \mathbb{P}(\text{fail=1}\mid E_1)+\frac{2}{T\Delta_i^2}+\frac{1}{\log^2 T},\end{aligned}\tag{89}$$

where the first and third inequalities are due to the law of total probability, the second inequality is due to (83), and the last inequality is due to (84). 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 10, we have

$$\begin{aligned}\mathbb{P}(\text{fail}=1\mid E_1)&\le\mathbb{P}\Bigg(\exists t_i\ge 1,\mu'-\theta_{i',t_i}+\sqrt{\frac{2}{t_i}\log\Big(\frac{T}{t_i}\Big(\log^2\frac{T}{t_i}+1\Big)\Big)}\le 0\mid E_1\Bigg)\\&\le\frac{4(16e^2+1)}{T(1-\epsilon_i)^2\Delta_i^2}.\end{aligned}\tag{90}$$

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

$$\exp(-\Delta_i^2\log^2 T/4)\le\frac{1}{T}.\tag{91}$$

Then, following the similar argument in (78), we can obtain

$$\mathbb{P}(\text{fail}=1,a=i)\le\frac{1}{T}.\tag{92}$$

Therefore, substituting (89), (90) and (92) into the definition of I_4 in (80), we have

$$\begin{aligned} I_4 &= \log^2 T\mathbb{P}(\text{fail} = 1) + T\mathbb{P}(\text{fail} = 1, a = i) \\ &\le 2 + \frac{2\log^2 T}{T\Delta_i^2} + \frac{4(16e^2+1)\log^2 T}{T(1-\epsilon_i)^2\Delta_i^2}. \end{aligned} \tag{93}$$

Completing the proof: we can choose a sufficiently large T such that all the conditions (82), (85), (91) are satisfied simultaneously. Substituting (93), (88), (87) and $I_1 \le 1$ back into (80), we have

$$\mathbb{E}[T_i] \le 4 + \frac{1}{(1-\epsilon_i)^2\Delta_i^2}\left(O(1) + 2\log(4T\Delta_i^2(\log^2(4T\Delta_i^2)+1)) + \sqrt{4\pi\log(4T\Delta_i^2(\log^2(4T\Delta_i^2)+1))}\right)$$

for all $i \ge 2$. Note that for $T \to \infty$, $\epsilon_i \le 1/(\log T)^{\frac{1}{3}}$. Hence we have $\lim_{T\to\infty}\mathbb{E}[T_i]/\log T = 2/\Delta_i^2$ and $\lim_{T\to\infty} R_\mu(T)/\log T = \sum_i 2/\Delta_i$. ■