Online Appendix for "Multi-armed Bandit Experimental Design: Online Decision-making and Adaptive Inference"

David Simchi-Levi and Chonghuan Wang

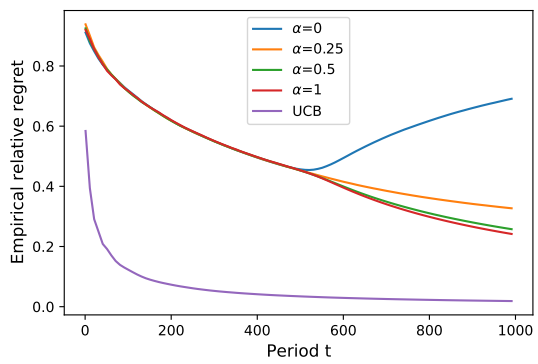
A. Numerical Results

In this section, we study the empirical performances of our algorithms on both the online decision-making efficiency and the ATE inference. We measure the performance of a learning algorithm π by the relative regret defined as $\frac{\sum_{t=1}^{n}\mu_{a^*}-\mu_{a_t}}{n\mu_{a^*}}$. We compare the quality of ATE inference with the traditional RCTs where at each time t, we uniformly randomly select an action and calculates ATE by the sample averages. For stochastic MAB experiments, we compare the learning efficiency with the well-known UCB algorithm. For each instance tested, we repeat the experiments for 1000 independent runs, and approximate the relative regret by the empirical relative regret averaged over the 1000 runs.

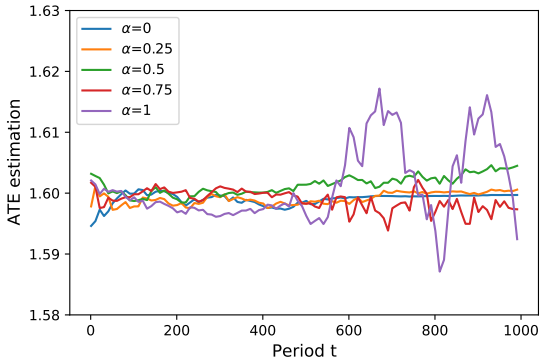
A.1. Numerical Results for Stochastic MAB Experiments

In this section, we numerically study the performance of our `EXP3E` and `EXP3EG`. We start from $K=2$ and the rewards of two arms are uniformly distributed in $[0.6,1]$ and $[-1,-0.6]$, respectively. In Figure 1(a), we compare the empirical relative regrets under different α. Since α begins to take effects only after the suboptimal arm is eliminated, the empirical relative regrets are almost the same at the beginning phase of the experiments. As shown in Figure 1(a), when α decreases, the regret increases and when $\alpha=0$, the regret grows almost linearly, which is consistent with our theory. Although UCB has the smallest regrets, it can not have statistical power guarantee as we mentioned before. In Figure 1(b), we show the estimated ATE at each period. When α grows, though the regret decreases, the ATE estimator becomes relatively unstable. From our theory, a larger α can only have weaker guarantee of the quality of ATE estimations. Even for $\alpha=1$, the largest error throughout the experiment is smaller than 1% shown in Figure 1(b). We also compare the ATE estimation with the RCT at the end of the experiments. Since we repeat each experiment 1000 times independently, we draw the box plot of the distribution of the 1000 estimations after the experiments in Figure EC.2, where the blue line is the average. It can be seen that RCT has the strongest statistical power, but it incurs a linear regret. When α increases, the mean values are close to the ATE, but the variance of the estimations grows large. This may also lead by the large variance issue of IPW estimators.

In Figure EC.3, we show the empirical relative regret for $K=3$, where the rewards are i.i.d. generated from the uniform distribution on $[0.6,1]$, $[-0.2,0.2]$, $[-1,-0.6]$ respectively. The observations are similar with those from $K=2$. UCB is the most efficient in terms of the empirical



(a) Comparison of empirical relative regret



(b) Comparison of ATE estimation

Figure EC.1 Empirical and adaptive ATE estimations for stochastic MAB experiments for $K=2$.

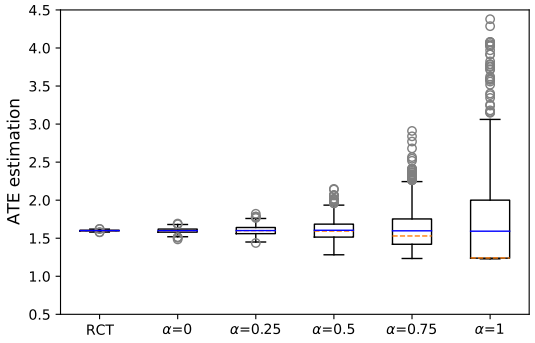


Figure EC.2 ATE estimations after experiments $K=2$.

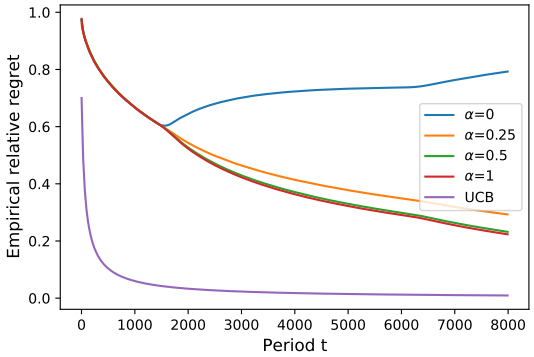
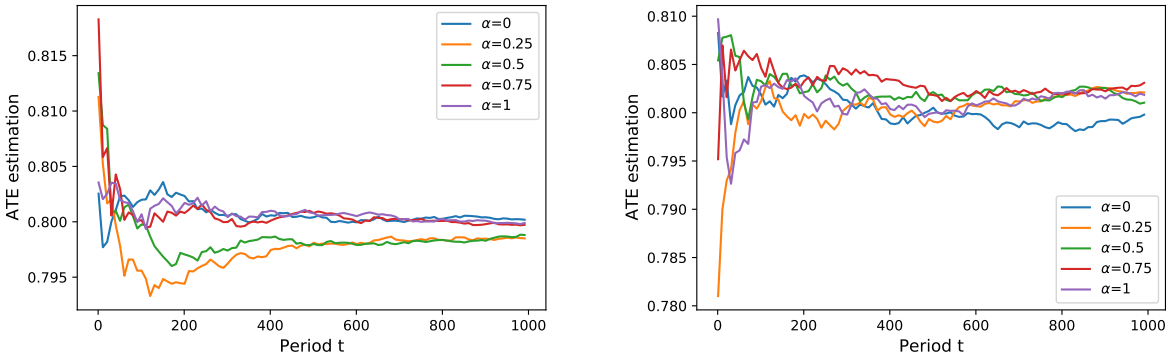


Figure EC.3 Empirical relative regret $K>2$.

relative regret. When α grows, the regret decreases. In Figure EC.4, we show the adaptive estimations of $\Delta^{(1,2)}$ and $\Delta^{(2,3)}$ of our algorithm through the whole experiments. Figure EC.5 shows the distribution of the $\hat{\Delta}^{(1,2)}$ and $\hat{\Delta}^{(2,3)}$ in 1000 independent experiments under different α and RCT. Figure 5(a) shows that the estimations of $\Delta^{(1,2)}$ are of better quality than the estimations of $\Delta^{(2,3)}$ empirically. It is because the third arm is always the first to be eliminated, and thus the observations of the first and the second arms are more sufficient. From Figure EC.5, the number of outliers grows with α, and the distribution of the outliers also becomes wider.

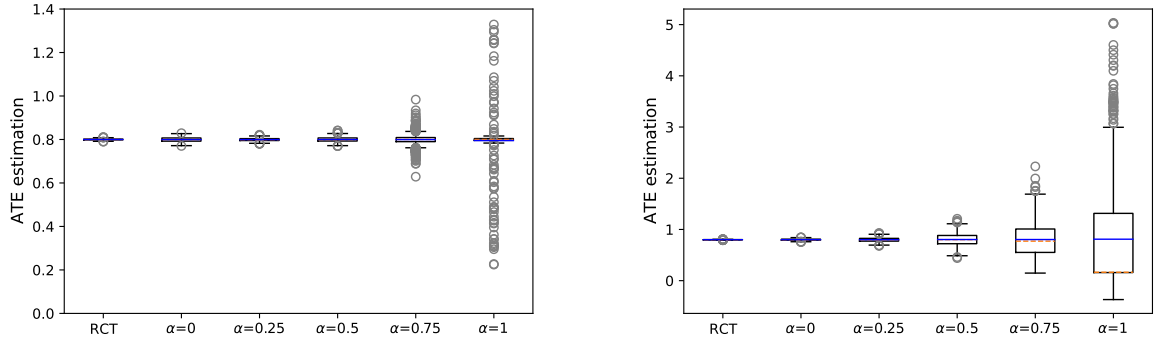
A.2. Numerical Results for MAB Experiments with Adversarial Baseline Rewards

In this section, we investigate the empirical results of our algorithm in Section 2 when the rewards have an adversarial baseline. We set $K=2$, $f_t=0.5\times sin(t)$, $\mu_1=0.8$, $\mu_2=-0.8$ and the noise to be uniformly distributed on $[-0.2,0.2]$. Figure 6(a) presents the empirical relative regrets, which are similar to the results in Figure 1(a). This shows the robustness of our algorithm. In Figure 6(b), we compare the adaptive ATE estimations under different α. Similarly, when α decreases,



(a) Adaptive estimations of $\Delta^{(1,2)}$ (b) Adaptive estimations of $\Delta^{(2,3)}$

Figure EC.4 Adaptive ATE estimations for stochastic MAB experiments for $K=3$.



(a) Estimations of $\Delta^{(1,2)}$ after experiments (b) Estimations of $\Delta^{(2,3)}$ after experiments

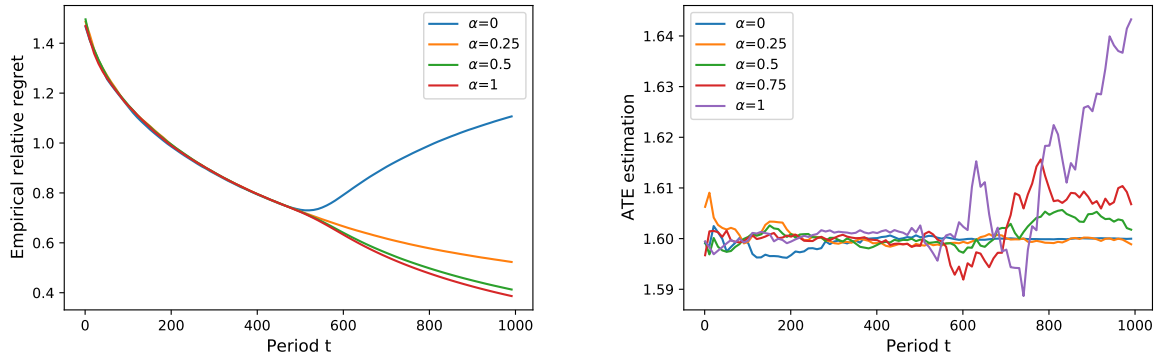
Figure EC.5 Distribution of ATE estimations for stochastic MAB experiments for $K=3$.

the ATE estimator becomes more stable. As shown in Figure 6(b), even with the adversarial baseline rewards, our estimator can still infer the ATE with high quality. Figure EC.7 presents the distribution of the ATE estimations at the end of the experiment over the 1000 independent simulations. Surprisingly, RCT still has sufficient statistical power. The possible reason may be that the design of f_t is too "simple". In real application, f_t may be very hard to model with a simple function.

B. Technical Details for Section 2

B.1. Proof to Lemma 1

Now, we provide the proof to Lemma 1. First, we start from proving Lemma 1. We define a Rademacher-like distribution as $X \sim Rad(p)$ means $X=-1$ with probability p and $X=1$ with probability $1-p$. Consider the following two bandits instance $\nu_1 = (Rad(\frac{1-\xi}{2}), Rad(\frac{1}{2}))$ and $\nu_2 = (Rad(\frac{1-\xi}{2}), Rad(\frac{1+2\phi(t)}{2}))$. Note that the treatment effects of ν_1 and ν_2 is $\Delta_1 = \xi$ and $\Delta_2 = \xi + 2\phi(t)$.



(a) Comparison of empirical relative regret (b) Comparison of ATE estimation

Figure EC.6 Empirical and adaptive ATE estimations for MAB experiments with adversarial baseline rewards.

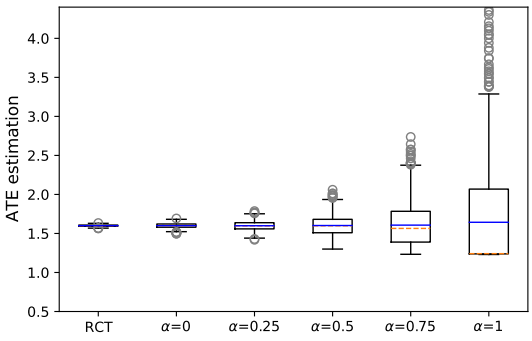


Figure EC.7 ATE estimations with adversarial baseline.

ξ can be any number in $(0,1]$. By such constructions and the symmetry, ν_1 and ν_2 can represent all the possible instances without loss of generality. We define the minimum distance test $\psi(\hat{\Delta}_t)$ that is associated to $\hat{\Delta}_t$ by $\psi(\hat{\Delta}_t) = \arg\min_{i=1,2} |\hat{\Delta}_t - \Delta_i|$. If $\psi(\hat{\Delta}_t) = 1$, we know that $|\hat{\Delta}_t - \Delta_1| \leq |\hat{\Delta}_t - \Delta_2|$. By the triangle inequality, we can have, if $\psi(\hat{\Delta}_t) = 1$,

$$|\hat{\Delta}_t - \Delta_2| \geq |\Delta_1 - \Delta_2| - |\hat{\Delta}_t - \Delta_1| \geq |\Delta_1 - \Delta_2| - |\hat{\Delta}_t - \Delta_2|,$$

which yields that $|\hat{\Delta}_t - \Delta_2| \geq \frac{1}{2}|\Delta_1 - \Delta_2| = \phi(t)$. Symmetrically, if $\psi(\hat{\Delta}_t) = 2$, we can have $|\hat{\Delta}_t - \Delta_1| \geq \frac{1}{2}|\Delta_1 - \Delta_2| = \phi(t)$. Therefore, we can use this to show

$$\begin{aligned}
\inf_{\hat{\Delta}_t} \max_{\nu \in \mathcal{E}_0} \mathbb{P}_\nu(|\hat{\Delta}_t - \Delta_\nu|_2 \geq \phi(t)) &\geq \inf_{\hat{\Delta}_t} \max_{i \in \{1,2\}} \mathbb{P}_{\nu_i}(|\hat{\Delta}_t - \Delta_i|_2 \geq \phi(t)) \\
&\geq \inf_{\hat{\Delta}_t} \max_{i \in \{1,2\}} \mathbb{P}_{\nu_i}(\psi(\hat{\Delta}_t) \neq i) \\
&\geq \inf_{\psi} \max_{i \in \{1,2\}} \mathbb{P}_{\nu_i}(\psi \neq i),
\end{aligned}$$

where the last infimum is taken over all tests ψ based on $\mathcal{H}_t$ that take values in $\{1,2\}$.

$$\begin{aligned}\inf_{\hat{\Delta}_t}\max_{\nu\in\mathcal{E}_0}\mathbb{P}_\nu(|\hat{\Delta}_t-\Delta_\nu|_2\geq\phi(t)) &\geq \inf_{\psi}\max_{i\in\{1,2\}}\mathbb{P}_{\nu_i}(\psi\neq i)\\ &\geq \frac{1}{2}\inf_{\psi}(\mathbb{P}_{\nu_1}(\psi=2)+\mathbb{P}_{\nu_2}(\psi=1))\\ &= \frac{1}{2}[1-\mathrm{TV}(\mathbb{P}_{\nu_1},\mathbb{P}_{\nu_2})]\\ &\geq \frac{1}{2}[1-\sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_{\nu_1},\mathbb{P}_{\nu_2})}]\\ &\geq \frac{1}{2}\left[1-\sqrt{\frac{8\phi(t)^2}{3\xi}\mathcal{R}_{\nu_1}(t,\pi)}\right],\end{aligned}$$

where the equality holds due to Neyman-Pearson lemma (see, Corollary EC.1 in Appendix C.2) and the third inequality holds due to Pinsker's inequality (see, Theorem EC.5 in Appendix C.3), and the fourth inequality holds due to the follows:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\nu_1},\mathbb{P}_{\nu_2}) &= \sum_{s=1}^{t}\mathbb{E}_{\nu_1}[\mathrm{KL}(P_{1,A_t},P_{2,A_t})] = \sum_{i=1}^{2}\mathbb{E}_{\nu_1}[T_i(n)]\mathrm{KL}(P_{1,i},P_{2,i})\\ &= \frac{(\phi(t))^2}{\frac{1}{4}-(\phi(t))^2}(\mathbb{E}_{\nu_1}[T_i(n)]) \leq \frac{16\phi(t)^2}{3\xi}\mathcal{R}_{\nu_1}(t,\pi),\end{aligned}$$

where we use $\mathrm{KL}(Rad(\frac{1}{2}),Rad(\frac{1+2\phi(t)}{2}))=\frac{\phi(t)^2}{1/4-\phi(t)^2}\leq\frac{16\phi(t)^2}{3}$, and the last inequality holds because the history $\mathcal{H}_t$ is generated by π and $\xi\mathbb{E}_{\nu_1}[T_i(n)]$ is just the expected regret of ν_1, which is just the definition of regret. We finish the proof of Lemma 1. Q.E.D.

B.2. Proof to Theorem 1

Based on Lemma 1, given policy π, and $\hat{\Delta}_n$, if $\phi(n)\leq\sqrt{\frac{3|\Delta_u|}{32\mathcal{R}_u(n,\pi)}}$ for some $u\in\mathcal{E}_0$,

$$\max_{\nu\in\mathcal{E}_0}\mathbb{E}[|\hat{\Delta}_n-\Delta_\nu|]\geq\phi(t)\max_{\nu\in\mathcal{E}_0}\mathbb{P}_\nu(|\hat{\Delta}_t-\Delta_\nu|_2\geq\phi(t))\geq\frac{\phi(n)}{2}\left[1-\sqrt{\frac{8\phi(t)^2}{3\xi}\mathcal{R}_{\nu_1}(t,\pi)}\right]\geq\frac{\phi(n)}{4},$$

where the second inequality holds due to Lemma 1. We use $\nu_{\pi,\hat{\Delta}_n}$ to denote $\arg\max_{\nu\in\mathcal{E}_0}\mathbb{E}[|\hat{\Delta}_n-\Delta_\nu|]$ given policy π and $\hat{\Delta}$, and thus $e_{\nu_{\pi,\hat{\Delta}_n}}(n,\hat{\Delta}_n)\geq\frac{\phi(n)}{4}$. After taking $\phi(n)=\sqrt{\frac{3|\Delta_{\nu_{\pi,\hat{\Delta}_n}}|}{32\mathcal{R}_{\nu_{\pi,\hat{\Delta}_n}}(n,\pi)}}$, we retrieve for any given policy π and $\hat{\Delta}_n$,

$$\max_{\nu\in\mathcal{E}_0}\left[e_\nu(n,\hat{\Delta}_n)\sqrt{\mathcal{R}_\nu(n,\pi)}\right]\geq e_{\nu_{\pi,\hat{\Delta}_n}}(n,\hat{\Delta}_n)\sqrt{\mathcal{R}_{\nu_{\pi,\hat{\Delta}_n}}(n,\pi)}\geq\frac{\phi(n)}{4}\sqrt{\mathcal{R}_{\nu_{\pi,\hat{\Delta}_n}}(n,\pi)}=\Theta(1),$$

where the last equation holds because we plug in $\phi(n)$ and $\Delta_\nu=\Theta(1)$ for $\nu\in\mathcal{E}_0$. Since the above inequalities hold for any policy π and $\hat{\Delta}_n$, we finish the proof. Q.E.D.

B.3. Proof to Theorem 2

We conduct proof by contradiction. Assume that $(\pi_0, \hat{\Delta}_0)$ satisfies Eq. (5), but is not Pareto optimal. This means that there exists a $(\pi_1, \hat{\Delta}_1)$ that Pareto dominates $(\pi_0, \hat{\Delta}_0)$. The lower bound in Eq. (4) guarantees that there must be a point at the front of $(\pi_1, \hat{\Delta}_1)$, denoted by $(e_{\nu_1}(n, \hat{\Delta}_1), \mathcal{R}_{\nu_1}(n, \pi_1))$ satisfying $e_{\nu_1}(n, \hat{\Delta}_1)\sqrt{\mathcal{R}_{\nu_1}(n, \pi_1)} = \Omega(1)$. By the definition of Pareto dominate, there exists $(e_{\nu_2}(n, \hat{\Delta}_0), \mathcal{R}_{\nu_2}(n, \pi_0)) \in \mathcal{F}(\pi_0, \hat{\Delta}_0)$ such that

$$e_{\nu_2}(n, \hat{\Delta}_0)\sqrt{\mathcal{R}_{\nu_2}(n, \pi_0)} > e_{\nu_1}(n, \hat{\Delta}_1)\sqrt{\mathcal{R}_{\nu_1}(n, \pi_1)} = \Omega(1).$$

Note that, as we have mentioned, the strict inequality in the above is in the term of the dependence of n. It means that $(e_{\nu_2}(n, \hat{\Delta}_0), \mathcal{R}_{\nu_2}(n, \pi_0)) = \Omega(n^p)$ for some strictly positive $p > 0$, which contradicts with our assumption that Eq. (5) holds. We finish the proof. Q.E.D.

B.4. Proof to Lemma 2

Note that the algorithm runs in a two-phase manner. In the first phase the main task is to identify the best arm, and in the second phase the main task is to gain information to conduct inference. Without loss of generality, we assume the first arm is optimal so that $\mu_1 = \mu_a^*$. Then, we can have $\mathcal{R}(n, \pi) = \Delta\mathbb{E}[T_2(n)]$. We define two stopping time $\tau(a) = \max\{t : a \in \mathcal{A}_t\}$ for $a \in \{1, 2\}$. Thus, the first stage ends after $\min_{a\in\{1,2\}} \tau(a)$.

For every period $t \le \min_{a\in\{1,2\}} \tau(a)$, we have $\hat{R}_{t-1}(a) \ge \hat{R}_{t-1}^{max} - 2\sqrt{C(t-1)}$ for $a \in \{1, 2\}$. Therefore, we can have

$$\frac{1}{\pi_t(a)} = \frac{e^{\varepsilon_{t-1}\hat{R}_{t-1}(1)} + e^{\varepsilon_{t-1}\hat{R}_{t-1}(2)}}{e^{\varepsilon_{t-1}\hat{R}_{t-1}(a)}} \le 1 + e^{\varepsilon_{t-1}(\hat{R}_{t-1}^{\max} - \hat{R}_{t-1}^{\max} + 2\sqrt{C(t-1)})} = 1 + e^2 := C'. \tag{EC.1}$$

Denote $M_t^1 = \hat{R}_t(1) - t\mu_1$ and $M_t^2 = \hat{R}_t(2) - t\mu_2$. Note that M_t^1 and M_t^2 are both martingales. Also, $M_{t\wedge\tau(1)}^1 := M_t^1 \mathbb{I}_{t\le\tau(1)}$ and $M_{t\wedge\tau(2)}^2 := M_t^2 \mathbb{I}_{t\le\tau(2)}$ are two stopped martingale. The variance of M_t^a can be written as

$$\begin{aligned}\sum_{t=1}^{n} \mathbb{E}\left[\left(\frac{R_t}{\pi_t(a)}\mathbb{I}_{a=A_t} - \mu_a\right)^2 \mid \mathcal{H}_{t-1}\right] &= \sum_{t=1}^{n} \mathbb{E}\left[\left(\frac{R_t}{\pi_t(a)}\mathbb{I}_{a=A_t}\right)^2 \mid \mathcal{H}_{t-1}\right] - t\mu_a^2 \\ &\le \sum_{t=1}^{n} \pi_t(a)\frac{R_t^2}{\pi_t^2(a)} \le \sum_{t=1}^{n} \frac{1}{\pi_t(a)},\end{aligned}$$

where we used $|R_t| \le 1$. We use $V_t(a) := \sum_{\tau=1}^{t} \frac{1}{\pi_\tau(a)}$ to denote the upper bound on the variance of the martingale M_t^a. By Eq. (EC.1), we can know that for any $t \le \min_{a\in\{1,2\}} \tau(a)$, $V_t(a) \le (C't) \vee \frac{(e^2+2)^2}{(e-2)} \ln(2/\delta) \le \frac{(e^2+2)^2}{(e-2)} \ln(2/\delta)t$ for any $\delta \le 2/e$. Therefore, by Bernstein's inequality, with probability at least $1-\delta$, we have, for all $t \le \min_{a\in\{1,2\}} \tau(a)$,

$$|M_t^a| \le 2\sqrt{(e^2+2)^2 t\left(\log\frac{2}{\delta}\right)^2} = \sqrt{Ct}. \tag{EC.2}$$

We first define a good event $\mathcal{E}_1 = \{\tau(1) = n\}$, which means that the optimal arm is not eliminated during the whole planning horizon. The probability of the inverse event of $\mathcal{E}_1$ can be bounded as:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1^c) &= \mathbb{P}(\tau(1) < n) \\
&= \mathbb{P}(\cup_{t=1}^{n-1}\{\tau(1) = t\}) \\
&= \sum_{t=1}^{n-1} \mathbb{P}(\tau(1) = t) \\
&= \sum_{t=1}^{n-1} \mathbb{P}\left(\tau(1) = t, \left\{\hat{R}_t(2) - \hat{R}_t(1) \geq 2\sqrt{Ct}\right\}\right) \\
&\leq \sum_{t=1}^{n-1} \mathbb{P}\left(\tau(1) = t, \left\{\hat{R}_t(1) \leq \mu_1 t - \sqrt{Ct}\right\}\right) + \mathbb{P}\left(\tau(1) = t, \left\{\hat{R}_t(2) \geq \mu_2 t + \sqrt{Ct}\right\}\right) \qquad \text{(EC.3)} \\
&= \sum_{t=1}^{n-1} \mathbb{P}\left(\tau(1) = t, M_t^1 \leq -\sqrt{Ct}\right) + \mathbb{P}\left(\tau(1) = t, M_t^2 \geq \sqrt{Ct}\right) \\
&\leq 2(n-1)\delta, \qquad \text{(EC.4)}
\end{aligned}$$

where the second equality holds since $\{\tau(1) = t\}$ are exclusive events and the last inequality holds due to Eq. (EC.2). Eq. (EC.3) follows from the fact that if $\left\{\hat{R}_t(2) - \hat{R}_t(1) \geq 2\sqrt{Ct}\right\}$ holds, then at least one of $\left\{\hat{R}_t(1) \leq \mu_1 t - \sqrt{Ct}\right\}$ and $\left\{\hat{R}_t(2) \geq \mu_2 t + \sqrt{Ct}\right\}$ happens. Otherwise, we can have $\hat{R}_t(2) - \hat{R}_t(1) \leq (\mu_2 - \mu_1)t + 2\sqrt{Ct} < 2\sqrt{Ct}$.

We finish the proof of Lemma 2. Q.E.D.

B.5. Proof to Lemma 3

Now, we elaborate on bounding the stopping time $\tau(2)$. We consider the following probability,

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_1, \tau(2) \geq \frac{16C}{\Delta^2} + 1\right) &= \sum_{t=\frac{16C}{\Delta^2}+1}^{n} \mathbb{P}\left(\mathcal{E}_1, \tau(2) = t\right) \\
&= \sum_{t=\frac{16C}{\Delta^2}+1}^{n} \mathbb{P}\left(\mathcal{E}_1, \tau(2) = t, \hat{R}_{t-1}(1) - \hat{R}_{t-1}(2) \leq 2\sqrt{C(t-1)}\right) \\
&\leq \sum_{t=\frac{16C}{\Delta^2}+1}^{n} \mathbb{P}\left(\mathcal{E}_1, \tau(2) = t, \hat{R}_{t-1}(1) \leq \mu_1(t-1) - \sqrt{C(t-1)}\right) \\
&\qquad + \mathbb{P}\left(\mathcal{E}_1, \tau(2) = t, \hat{R}_{t-1}(2) \geq \mu_2(t-1) + \sqrt{C(t-1)}\right) \qquad \text{(EC.5)} \\
&\leq 2\left(n - \frac{16C}{\Delta^2}\right)\delta, \qquad \text{(EC.6)}
\end{aligned}$$

where the first equality holds because $\{\tau(2) = t\}$ are exclusive events, the second equality holds since $\mathcal{E}_1$ and $\tau(2) = t$ indicate that $\hat{R}_{t-1}(1) - \hat{R}_{t-1}(2) \leq 2\sqrt{C(t-1)}$ due to our elimination rule, and the last inequality holds due to Eq. (EC.2). The reason why Eq. (EC.5) holds is as following.

If $\hat{R}_{t-1}(1) > \mu_1(t-1) - \sqrt{C(t-1)}$ and $\hat{R}_{t-1}(2) < \mu_2(t-1) + \sqrt{C(t-1)}$ occurs at the same time, we can have

$$\begin{aligned}\hat{R}_{t-1}(1) - \hat{R}_{t-1}(2) &> (\mu_1 - \mu_2)(t-1) - 2\sqrt{C(t-1)} = \Delta(t-1) - 2\sqrt{C(t-1)} \\ &\geq \sqrt{\frac{16C}{t-1}}(t-1) - 2\sqrt{C(t-1)} = 2\sqrt{C(t-1)},\end{aligned}$$

where the second inequality holds because $t \geq \frac{16C}{\Delta^2} + 1$ tells that $\Delta \geq \sqrt{\frac{16C}{t-1}}$.

We finish the proof of Lemma 2. Q.E.D.

B.6. Proof to Theorem 3

Based on Lemmas 2 and 3, we can bound $\mathbb{E}[T_2(n)]$ as following

$$\begin{aligned}\mathbb{E}[T_2(n)] &= \mathbb{E}[T_2(n)\mathbb{I}_{\mathcal{E}_1}] + \mathbb{E}[T_2(n)\mathbb{I}_{\mathcal{E}_1^c}] \\ &\leq \mathbb{E}[\tau(2)\mathbb{I}_{\mathcal{E}_1}] + \mathbb{E}\left[\left(\sum_{t=\tau(2)+1}^{n} \mathbb{I}_{a_t=2}\right)\mathbb{I}_{\mathcal{E}_1}\right] + n\mathbb{P}(\mathcal{E}_1^c) \\ &\leq \left(\frac{16C}{\Delta^2} + 1\right) + \sum_{t=1}^{n} \frac{1}{t^\alpha} + n\mathbb{P}\left(\mathcal{E}_1, \tau(2) \geq \frac{16C}{\Delta^2} + 1\right) + n\mathbb{P}(\mathcal{E}_1^c) \\ &\leq \frac{16C}{\Delta^2} + \left(\frac{n^{1-\alpha}}{1-\alpha}\right) \wedge \log(n) + 1 + 4n^2\delta.\end{aligned}$$

Thus, the total regret can be bounded as $\mathcal{O}\left(\frac{\log n}{\Delta} + n^{1-\alpha}\log n\Delta\right)$. We finish the proof. Q.E.D.

B.7. Proof to Theorem 4

Based on the martingale we defined in Section 2.2, $M_t^{(1,2)} = M_t^1 - M_t^2$. We first bound the difference between $M_t^{(1,2)}$ and $M_{t-1}^{(1,2)}$ as

$$\left|M_t^{(1,2)} - M_{t-1}^{(1,2)}\right| = \left|\frac{R_t}{\pi_t(1)}\mathbb{I}_{A_t=1} - \frac{R_t}{\pi_t(2)}\mathbb{I}_{A_t=2} - \Delta\right| \leq 2 + \frac{1}{\pi_t(1)} + \frac{1}{\pi_t(2)}.$$

The variance of the martingale V_t^0 can be bounded as

$$\begin{aligned}\sum_{t=1}^{n}\mathbb{E}\left[\left(\frac{R_t}{\pi_t(1)}\mathbb{I}_{A_t=1} - \frac{R_t}{\pi_t(2)}\mathbb{I}_{A_t=2} - \Delta\right)^2 \mid \mathcal{H}_{t-1}\right] &= \sum_{t=1}^{n}\mathbb{E}\left[\left(\frac{R_t}{\pi_t(1)}\mathbb{I}_{A_t=1} - \frac{R_t}{\pi_t(2)}\mathbb{I}_{A_t=2}\right)^2 \mid \mathcal{H}_{t-1}\right] - t\Delta^2 \\ &\leq \sum_{t=1}^{n}\frac{1}{\pi_t(1)} + \frac{1}{\pi_t(2)}.\end{aligned}$$

Thus, by Eq. (EC.1), $V_t^0 \leq 2C't$ when t is in the first phase. When t falls into the second stage, we can have $V_t^0 \leq 2C'(\tau(1)\wedge\tau(2)) + 2\frac{(t+1)^{1+\alpha}}{1+\alpha} \leq 2C't + 2\frac{(t+1)^{1+\alpha}}{1+\alpha} \leq (4+2e^2)(t+1)^{1+\alpha}$. In order to use Bernstein's inequality again, we need to further control V_t^0 to satisfy the condition in Theorem EC.3 as following. For any $\delta \leq 2/e$,

$$V_t^0 \leq \left((4+2e^2)(t+1)^{1+\alpha}\right) \vee \left[\frac{(4+2e^2)^2 \vee (4+2t^\alpha)^2}{e-2}\ln(\frac{2}{\delta})\right]$$

$$\leq \left((4+2e^2)(t+1)^{1+\alpha}\right) \vee \left[\frac{(4+2e^2)^2 \vee 36(t+1)^{1+\alpha}}{e-2}\ln(\frac{2}{\delta})\right]$$
$$\leq \left((4+2e^2)(t+1)^{1+\alpha}\right) \vee \left[\frac{(4+2e^2)^2(t+1)^{1+\alpha}}{e-2}\ln(\frac{2}{\delta})\right] \leq \frac{(4+2e^2)^2(t+1)^{1+\alpha}}{e-2}\ln(\frac{2}{\delta}),$$

where the first inequality is trivial since we just put in an extra term and take the maximum, the second inequality holds due to $(4+2t^\alpha)^2 \leq 36(t+1)^{1+\alpha}$ for $\alpha \in [0,1]$, the third inequality is just because of $(4+2e^2)^2 > 36$.

By Bernstein's inequality, we have, with probability $1-\delta$ for all t:

$$\left|M_t^{(1,2)}\right| \leq (4e^2+8)\log\frac{2}{\delta}\sqrt{(t+1)^{1+\alpha}} \leq (8e^2+16)\log\frac{2}{\delta}\sqrt{t^{1+\alpha}},$$

where the second inequality holds because $(t+1)^{1+\alpha} = t^{1+\alpha}(1+\frac{1}{t})^{1+\alpha} \leq 4t^{1+\alpha}$. By dividing both sides of t, we can have

$$\left|\hat{\Delta}_t - \Delta\right| \leq \frac{(8e^2+16)\log\frac{2}{\delta}}{\sqrt{t^{1-\alpha}}}.$$

We finish the proof. Q.E.D.

B.8. Proof to Theorem 5

We again conduct the proof by contradiction. Suppose there is a Pareto optimal policy $(\pi_p, \hat{\Delta}_p)$ with an instance ν_p such that $e_{\nu_p}(n, \hat{\Delta}_p) \simeq n^{-\beta}$ and $\mathcal{R}_{\nu_p}(n, \pi_p) \simeq n^{2\gamma}$ for some positive β, $\gamma \leq \frac{1}{2}$ and $\gamma - \beta > 0$. As befrore, here we ignore the constant and the logarithm factors. By taking $\alpha = 1-2\gamma$ in Algorithm 1, the front of $(\pi_{1-2\gamma}, \hat{\Delta}_{1-2\gamma})$ is $\{(n^{-\gamma}, n^{2\gamma})\}$. Since $\gamma > \beta$, $(\pi_{1-2\gamma}, \hat{\Delta}_{1-2\gamma})$ Pareto dominates $(\pi_p, \hat{\Delta}_p)$, which contradicts with the assumption that $(\pi_p, \hat{\Delta}_p)$ is Pareto optimal. We finish the proof. Q.E.D.

C. Technical Details for Section 3

C.1. Proof to Theorem 6

For each arm a, there are two phases in our algorithm. The first phase is that $a \in \mathcal{A}_t$, i.e., a is not eliminated. The second phase is that we force the policy to explore a even if it has been ruled out to be the optimal choice. Without loss of generality, we assume the first arm is optimal so that $\mu_1 = \mu_a^*$. Then, we can have $\mathcal{R}(n,\pi) = \sum_{a\in\{2,\cdots,K\}} \Delta(a)\mathbb{E}[T_a(n)]$, where $T_a(n)$ is the random variable of the number of the times that arm action a is played.

For period $t \geq 2$, we consider $a \in \mathcal{A}_t$ and have $\hat{R}_{t-1}(a) \geq \hat{R}_{t-1}^{max} - 2\sqrt{C(t-1)}$. Therefore, we can have

$$\frac{1}{\pi_t(a)} = \frac{1}{(1-|\mathcal{A}_t^c|\alpha_t)}\frac{\sum_{a'\in\mathcal{A}_t} e^{\varepsilon_{t-1}\hat{R}_{t-1}(a')}}{e^{\varepsilon_{t-1}\hat{R}_{t-1}(a)}} \leq \frac{1}{(1-\frac{K-1}{Kt^\alpha})}\frac{\sum_{a'\in\mathcal{A}_t} e^{\varepsilon_{t-1}\hat{R}_{t-1}(a')}}{e^{\varepsilon_{t-1}\hat{R}_{t-1}(a)}}$$
$$\leq K|\mathcal{A}_t|(1+e^{\varepsilon_{t-1}(\hat{R}_{t-1}^{\max}-\hat{R}_{t-1}^{\max}+2\sqrt{C(t-1)})}) \leq K^2(1+e^2) := C'. \tag{EC.7}$$

Denote $M_t^a = \hat{R}_t(a) - t\mu_a$. Note that M_t^a is a martingales for all $a \in [K]$ and $t \in [n]$. We also define the stopping time for each arm a as $\tau(a) = \max\{t : a \in \mathcal{A}_t\}$. The variance of M_t^a can be written as

$$\sum_{t=1}^{n} \mathbb{E}\left[\left(\frac{R_t}{\pi_t(a)}\mathbb{I}_{a=A_t} - \mu_a\right)^2 \mid \mathcal{H}_{t-1}\right] = \sum_{t=1}^{n} \mathbb{E}\left[\left(\frac{R_t}{\pi_t(a)}\mathbb{I}_{a=A_t}\right)^2 \mid \mathcal{H}_{t-1}\right] - t\mu_a^2$$
$$\leq \sum_{t=1}^{n} \pi_t(a)\frac{1}{\pi_t^2(a)} \leq \sum_{t=1}^{n} \frac{1}{\pi_t(a)},$$

where we used $|R_t| \leq 1$. We use $V_t(a) := \sum_{\tau=1}^{t} \frac{1}{\pi_\tau(a)}$ to denote the upper bound on the variance of the martingale M_t^a. By Eq. (EC.7), we can know that for any $t \leq \tau(a)$, $V_t(a) \leq C't \vee \frac{(C'+1)^2}{e-2}\ln(2/\delta) \leq \frac{(C'+1)^2}{e-2}\ln(2/\delta)t$. Therefore, by Bernstein's inequality, with probability at least $1-\delta$, we have, for all $t \leq \tau(a)$,

$$|M_t^a| \leq 2\sqrt{(C'+1)^2\left(\log\frac{2}{\delta}\right)^2 t} \leq \sqrt{Ct}. \tag{EC.8}$$

Similarly, we first define a good event $\mathcal{E}_1 = \{\tau(1) = n\}$, which tells that the optimal arm is not eliminated. The probability of the inverse event of $\mathcal{E}_1$ can be bounded as:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1^c) &= \mathbb{P}(\tau(1) < n) \\
&= \mathbb{P}(\cup_{t=1}^{n-1}\{\tau(1) = t\}) \\
&= \sum_{t=1}^{n-1} \mathbb{P}(\tau(1) = t) \\
&= \sum_{t=1}^{n-1} \mathbb{P}\left(\tau(1) = t, \left\{\exists a \in \mathcal{A}_t, \hat{R}_t(a) - \hat{R}_t(1) \geq 2\sqrt{Ct}\right\}\right) \\
&= \sum_{t=1}^{n-1} \mathbb{P}\left(\tau(1) = t, \left\{\exists a \in \mathcal{A}_t, \tau(a) > t \text{ and } \hat{R}_t(a) - \hat{R}_t(1) \geq 2\sqrt{Ct}\right\}\right) \\
&\leq \sum_{t=1}^{n-1}\sum_{a=2}^{K} \mathbb{P}\left(\tau(1) = t, \tau(a) > t, \hat{R}_t(a) - \hat{R}_t(1) \geq 2\sqrt{Ct}\right) \\
&\leq \sum_{t=1}^{n-1}\sum_{a=2}^{K} \mathbb{P}\left(\tau(1) = t, \tau(a) > t, \hat{R}_t(1) \leq \mu_1 t - \sqrt{Ct}\right) \\
&\qquad + \mathbb{P}\left(\tau(1) = t, \tau(a) > t, \hat{R}_t(a) \geq \mu_a t + \sqrt{Ct}\right) \qquad \text{(EC.9)} \\
&= \sum_{t=1}^{n-1}\sum_{a=2}^{K} \mathbb{P}\left(\tau(1) = t, \tau(a) > t, M_t^1 \leq -\sqrt{Ct}\right) + \mathbb{P}\left(\tau(1) = t, \tau(a) > t, M_t^a \geq \sqrt{Ct}\right) \\
&\leq 2(n-1)(K-1)\delta, \qquad \text{(EC.10)}
\end{aligned}$$

where the third equality holds since $\{\tau(1) = t\}$ are exclusive events, the fourth equality follows from our elimination rule, the fifth equality is because of $a \in \mathcal{A}_t$, the first inequality is due to the union bound, and the last inequality holds due to Eq. (EC.8). Eq. (EC.9) follows from

the fact that if $\left\{\hat{R}_t(a)-\hat{R}_t(1)\geq 2\sqrt{Ct}\right\}$ holds, then at least one of $\left\{\hat{R}_t(1)\leq \mu_1 t-\sqrt{Ct}\right\}$ and $\left\{\hat{R}_t(a)\geq \mu_a t+\sqrt{Ct}\right\}$ happens. Otherwise, we can have $\hat{R}_t(a)-\hat{R}_t(1)\leq(\mu_a-\mu_1)t+2\sqrt{Ct}<2\sqrt{Ct}$.

Now, we elaborate on bounding the stopping time $\tau(a)$. We consider the following probability,

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_1,\tau(a)\geq\frac{16C}{\Delta(a)^2}+1\right)&=\sum_{t=\frac{16C}{\Delta(a)^2}+1}^{n}\mathbb{P}\left(\mathcal{E}_1,\tau(a)=t\right)\\
&=\sum_{t=\frac{16C}{\Delta(a)^2}+1}^{n}\mathbb{P}\left(\mathcal{E}_1,\tau(a)=t,\hat{R}_{t-1}(1)-\hat{R}_{t-1}(a)\leq 2\sqrt{C(t-1)}\right)\\
&\leq\sum_{t=\frac{16C}{\Delta(a)^2}+1}^{n}\mathbb{P}\left(\mathcal{E}_1,\tau(a)=t,\hat{R}_{t-1}(1)\leq\mu_1(t-1)-\sqrt{C(t-1)}\right)\\
&\qquad+\mathbb{P}\left(\mathcal{E}_1,\tau(a)=t,\hat{R}_{t-1}(a)\geq\mu_a(t-1)+\sqrt{C(t-1)}\right)\\
&\leq 2\left(n-\frac{16C}{\Delta(a)^2}\right)\delta,
\end{aligned}$$

where the first equality holds because $\{\tau(2)=t\}$ are exclusive events, the second equality holds since $\mathcal{E}_1$ and $\tau(a)=t$ indicate that $\hat{R}_{t-1}(1)-\hat{R}_{t-1}(a)\leq 2\sqrt{C(t-1)}$ due to our elimination rule, and the last inequality holds due to Eq. (EC.8). The reason why Eq. (EC.5) holds is as following. If $\hat{R}_{t-1}(1)>\mu_1(t-1)-\sqrt{C(t-1)}$ and $\hat{R}_{t-1}(a)<\mu_a(t-1)+\sqrt{C(t-1)}$ occurs at the same time, we can have

$$\begin{aligned}
\hat{R}_{t-1}(1)-\hat{R}_{t-1}(a)&>(\mu_1-\mu_a)(t-1)-2\sqrt{C(t-1)}=\Delta(a)(t-1)-2\sqrt{C(t-1)}\\
&\geq\sqrt{\frac{16C}{t-1}}(t-1)-2\sqrt{C(t-1)}=2\sqrt{C(t-1)},
\end{aligned}$$

where the second inequality holds because $t\geq\frac{16C}{\Delta(a)^2}+1$ tells that $\Delta(a)\geq\sqrt{\frac{16C}{t-1}}$.

Therefore, we can bound $\mathbb{E}[T_a(n)]$ as following

$$\begin{aligned}
\mathbb{E}[T_a(n)]&=\mathbb{E}[T_a(n)\mathbb{I}_{\mathcal{E}_1}]+\mathbb{E}[T_a(n)\mathbb{I}_{\mathcal{E}_1^c}]\\
&\leq\mathbb{E}[\tau(a)\mathbb{I}_{\mathcal{E}_1}]+\mathbb{E}\left[\left(\sum_{t=\tau(a)+1}^{n}\mathbb{I}_{A_t=a}\right)\mathbb{I}_{\mathcal{E}_1}\right]+n\mathbb{P}(\mathcal{E}_1^c)\\
&\leq\left(\frac{16C}{\Delta(a)^2}+1\right)+\sum_{t=1}^{n}\frac{1}{Kt^{\alpha}}+n\mathbb{P}\left(\mathcal{E}_1,\tau(a)\geq\frac{16C}{\Delta(a)^2}+1\right)+n\mathbb{P}(\mathcal{E}_1^c)\\
&\leq\frac{16C}{\Delta(a)^2}+\left(\frac{n^{1-\alpha}}{K(1-\alpha)}\right)\wedge\frac{\log(n)}{K}+1+2Kn^2\delta.
\end{aligned}$$

Thus, the total regret can be bounded as $\mathcal{O}(\sum_{a\in[K]/\{a^*\}}\frac{\log(n)}{\Delta(a)}+\Delta(a)n^{1-\alpha}\log(n))$. Q.E.D.

C.2. Proof to Theorem 7

We first bound the difference between $M_t^{(i,j)}$ and $M_{t-1}^{(i,j)}$ as

$$\left|M_t^{(i,j)} - M_{t-1}^{(i,j)}\right| = \left|\frac{R_t}{\pi_t(i)}\mathbb{I}_{A_t=i} - \frac{R_t}{\pi_t(j)}\mathbb{I}_{A_t=j} - (\mu_i - \mu_j)\right| \le 2 + \frac{1}{\pi_t(i)} + \frac{1}{\pi_t(j)}.$$

The variance of the martingale $V_t^{(i,j)}$ can be bounded as

$$\begin{aligned}
&\sum_{t=1}^{n} \mathbb{E}\left[\left(\frac{R_t}{\pi_t(i)}\mathbb{I}_{A_t=i} - \frac{R_t}{\pi_t(j)}\mathbb{I}_{A_t=j} - \Delta^{(i,j)}\right)^2 \mid \mathcal{H}_{t-1}\right] \\
&= \sum_{t=1}^{n} \mathbb{E}\left[\left(\frac{R_t}{\pi_t(i)}\mathbb{I}_{A_t=i} - \frac{R_t}{\pi_t(j)}\mathbb{I}_{A_t=j}\right)^2 \mid \mathcal{H}_{t-1}\right] - t(\Delta^{(i,j)})^2 \\
&\le \sum_{t=1}^{n} \frac{1}{\pi_t(i)} + \frac{1}{\pi_t(j)}. \qquad \text{(EC.11)}
\end{aligned}$$

We can have for all t, $V_t^{(i,j)} \le (C'\tau(i) + C'\tau(j)) + 2\frac{(t+1)^{1+\alpha}}{K(1+\alpha)} \le 2C't + 2\frac{(t+1)^{1+\alpha}}{K(1+\alpha)} \le (2C'+1)(t+1)^{1+\alpha}$. In order to apply Bernstein's inequality, we further control $V_t^{(i,j)}$ as

$$\begin{aligned}
V_t^{(i,j)} &\le (2C'+1)(t+1)^{1+\alpha} \vee \left[\frac{(2+2C')^2 \vee (2+2Kt^\alpha)^2}{(e-2)} \log\frac{2}{\delta}\right] \\
&\le (2C'+1)(t+1)^{1+\alpha} \vee \left[\frac{(2+2C')^2 \vee ((4+4K)^2(t+1)^{1+\alpha})}{(e-2)} \log\frac{2}{\delta}\right] \\
&\le \frac{(2+2C')^2(t+1)^{1+\alpha}}{(e-2)} \log\frac{2}{\delta}. \qquad \text{(EC.12)}
\end{aligned}$$

By Bernstein's inequality, we have, with probability $1-\delta$ for all t:

$$\left|M_t^{(i,j)}\right| \le 4(2+2C')\log\frac{2}{\delta}\sqrt{t^{1+\alpha}}.$$

By dividing both sides of t, we can have

$$\left|\hat{\Delta}_t^{(i,j)} - \Delta^{(i,j)}\right| \le \frac{4(2+2K^2(1+e^2))\log\frac{2}{\delta}}{\sqrt{t^{1-\alpha}}}.$$

Q.E.D.

D. Technical Details for Section 4

In this section, we will first prove Theorem 11, and then prove Theorem 10, for the simplicity in presentation.

D.1. Proof to Theorem 11

Without loss of generality, we assume the first arm is optimal so that $\mu_1 = \mu_a^*$. Then, we can have $\mathcal{R}(n,\pi) = \sum_{a\in\{2,\cdots,K\}} \Delta(a)\mathbb{E}[T_a(n)]$, where $T_a(n)$ is the random variable of the number of the times that arm action a is played. Define $\tau(a) := \max\{t : a \in \mathcal{A}_t\}$.

For every period t, we consider $a \in \mathcal{A}_t$ and Eq. (EC.7) still holds here i.e., $\frac{1}{\pi_t(a)} \le C'$.

Denote $M_t^{(i,j)} = \hat{R}_t(i) - \hat{R}_t(i) - t\Delta^{(i,j)}$. Note that $M_t^{(i,j)}$ is a martingale. Same as Eq. (EC.11), the variance of this martingale can be written as

$$\sum_{t=1}^{n} \mathbb{E}\left[\left(\frac{R_t}{\pi_t(i)}\mathbb{I}_{A_t=i} - \frac{R_t}{\pi_t(j)}\mathbb{I}_{A_t=j} - \Delta^{(i,j)}\right)^2 \mid \mathcal{H}_{t-1}\right] \le \sum_{t=1}^{n} \frac{1}{\pi_t(i)} + \frac{1}{\pi_t(j)}.$$

where we used $|R_t| \le 1$. We use $V_t^{(i,j)} := \sum_{\tau=1}^{t} \frac{1}{\pi_\tau(i)} + \frac{1}{\pi_\tau(j)}$ to denote the upper bound on the variance of the martingale $M_t^{(i,j)}$. By Eq. (EC.7), we can know that for any t and $i, j \in \mathcal{A}_t$, $V_t^{(i,j)} \le 2C't \vee \frac{4C'^2 \ln(2/\delta)}{e-2} \le \frac{4C'^2 \ln(2/\delta)}{e-2} t$. Therefore, by Bernstein's inequality, with probability at least $1-\delta$, we have, for any $t \le \tau(i) \wedge \tau(j)$,

$$|M_t^{(i,j)}| \le 2\sqrt{4C'^2 t \left(\log\frac{2}{\delta}\right)^2} \le 2\sqrt{Ct}. \tag{EC.13}$$

Similarly, we first define a good event $\mathcal{E}_1 = \{\tau(1) = n\}$, which tells that the optimal arm is not eliminated. The probability of the inverse event of $\mathcal{E}_1$ can be bounded as:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1^c) &= \mathbb{P}(\tau(1) < n) \\
&= \mathbb{P}(\cup_{t=1}^{n-1}\{\tau(1) = t\}) \\
&= \sum_{t=1}^{n-1} \mathbb{P}(\tau(1) = t) \\
&= \sum_{t=1}^{n-1} \mathbb{P}\left(\tau(1) = t, \left\{\exists a \in \mathcal{A}_t, \hat{R}_t(a) - \hat{R}_t(1) \ge 2\sqrt{Ct}\right\}\right) \\
&= \sum_{t=1}^{n-1} \mathbb{P}\left(\tau(1) = t, \left\{\exists a \in \mathcal{A}_t, \tau(a) > t \text{ and } \hat{R}_t(a) - \hat{R}_t(1) \ge 2\sqrt{Ct}\right\}\right) \\
&\le \sum_{t=1}^{n-1}\sum_{a=2}^{K} \mathbb{P}\left(\tau(1) = t, \tau(a) > t \text{ and } \hat{R}_t(a) - \hat{R}_t(1) \ge 2\sqrt{Ct}\right) \\
&\le \sum_{t=1}^{n-1}\sum_{a=2}^{K} \mathbb{P}\left(\tau(1) = t, \tau(a) > t \text{ and } \hat{R}_t(a) - \hat{R}_t(1) - t\Delta^{(a,1)} \ge 2\sqrt{Ct}\right) \\
&= \sum_{t=1}^{n-1}\sum_{a=2}^{K} \mathbb{P}\left(\tau(1) = t, \tau(a) > t \text{ and } M_t^{(a,1)} \ge 2\sqrt{Ct}\right) \\
&\le 2(n-1)(K-1)\delta,
\end{aligned}$$

where the third equality holds since $\{\tau(1) = t\}$ are exclusive events, the fourth equality follows from our elimination rule, the fifth equality is because of $a \in \mathcal{A}_t$, the first inequality is due to the union bound, and the last inequality holds due to Eq. (EC.13). The second inequality follows from $\Delta^{(a,1)} < 0$.

Now, we elaborate on bounding the stopping time $\tau(a)$. We consider the following probability,

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_1, \tau(a) \geq \frac{16C}{\Delta(a)^2}+1\right) &= \sum_{t=\frac{16C}{\Delta(a)^2}+1}^{n} \mathbb{P}\left(\mathcal{E}_1, \tau(a)=t\right) \\
&= \sum_{t=\frac{16C}{\Delta(a)^2}+1}^{n} \mathbb{P}\left(\mathcal{E}_1, \tau(a)=t, \hat{R}_{t-1}(1)-\hat{R}_{t-1}(a) \leq 2\sqrt{C(t-1)}\right) \\
&= \sum_{t=\frac{16C}{\Delta(a)^2}+1}^{n} \mathbb{P}\left(\mathcal{E}_1, \tau(a)=t, M_{t-1}^{(1,a)} \leq 2\sqrt{C(t-1)}-\Delta(a)(t-1)\right) \\
&\leq \sum_{t=\frac{16C}{\Delta(a)^2}+1}^{n} \mathbb{P}\left(\mathcal{E}_1, \tau(a)=t, M_{t-1}^{(1,a)} \leq -2\sqrt{C(t-1)}\right) \\
&\leq 2\left(n-\frac{16C}{\Delta(a)^2}\right)\delta.
\end{aligned}$$

where the first equality holds because $\{\tau(2)=t\}$ are exclusive events, the second equality holds since $\mathcal{E}_1$ and $\tau(a)=t$ indicate that $\hat{R}_{t-1}(1)-\hat{R}_{t-1}(a) \leq 2\sqrt{C(t-1)}$ due to our elimination rule, and the last inequality holds due to Eq. (EC.13). The first inequality holds because $t \geq \frac{16C}{\Delta(a)^2}+1$ tells that $\Delta(a) \geq \sqrt{\frac{16C}{t-1}}$.

Therefore, we can bound $\mathbb{E}[T_a(n)]$ as following

$$\begin{aligned}
\mathbb{E}[T_a(n)] &= \mathbb{E}[T_a(n)\mathbb{I}_{\mathcal{E}_1}] + \mathbb{E}[T_a(n)\mathbb{I}_{\mathcal{E}_1^c}] \\
&\leq \mathbb{E}[\tau(a)\mathbb{I}_{\mathcal{E}_1}] + \mathbb{E}\left[\left(\sum_{t=\tau(a)+1}^{n} \mathbb{I}_{A_t=a}\right)\mathbb{I}_{\mathcal{E}_1}\right] + n\mathbb{P}(\mathcal{E}_1^c) \\
&\leq \left(\frac{16C}{\Delta(a)^2}+1\right) + \sum_{t=1}^{n}\frac{1}{Kt^{\alpha}} + n\mathbb{P}\left(\mathcal{E}_1, \tau(a) \geq \frac{16C}{\Delta(a)^2}+1\right) + n\mathbb{P}(\mathcal{E}_1^c) \\
&\leq \frac{16C}{\Delta(a)^2} + \left(\frac{n^{1-\alpha}}{K(1-\alpha)}\right) \wedge \frac{\log(n)}{K} + 1 + 2Kn^2\delta.
\end{aligned}$$

Thus, the total regret can be bounded as $\mathcal{O}(\sum_{a\in[K]/\{a^*\}} \frac{\log(n)}{\Delta(a)} + \Delta(a)n^{1-\alpha}\log(n))$. Q.E.D.

D.2. Proof to Theorem 10

Similar as Eq. (EC.12), we can have $V_t^{(i,j)} \leq \frac{(2+2C')^2(t+1)^{1+\alpha}}{(e-2)}\log\frac{2}{\delta}$.

By Bernstein's inequality, we have, with probability $1-\delta$ for all t:

$$\left|M_t^{(i,j)}\right| \leq 4(2+2C')\log\frac{2}{\delta}\sqrt{t^{1+\alpha}}.$$

By dividing both sides of t, we can have

$$\left|\hat{\Delta}_t^{(i,j)} - \Delta^{(i,j)}\right| \leq \frac{4(2+2K^2(1+e^2))\log\frac{2}{\delta}}{\sqrt{t^{1-\alpha}}}.$$

Q.E.D.

D.3. Technical Details for Remark 2

We first formally restate our setting for the fully adversarial setting. During each time step $t \le n$, the environment generates a reward $r_t(a) \in [-1,1]$ for each arm $a \in \mathcal{A}$ based on a distribution $P_{a,t}$, where we allow $P_{a,t}$ to be adversarially chosen for different t. Accordingly, we define $\mu_{a,t} := \mathbb{E}[r_t(a)]$ and $\Delta_t^{(i,j)} = \mu_{i,t} - \mu_{j,t}$. Subsequently, the expected average outcome of arm a over time can be updated as $\mu_a = \frac{1}{n}\sum_{t=1}^n \mu_{a,t}$, and our estimand becomes $\Delta^{(i,j)} = \mu_i - \mu_j$ accordingly. Although we allow the existence of the adversary, we assume the best arm to remain the same over time, denoted as a^*. Note that we are not asking the rank of the other arms to remain fixed. We additionally introduce the notation $\Delta_t(a) = \mu_{a^*,t} - \mu_{a,t}$. Also, we still need the large-gap assumption, i.e., $\Delta_t(a) = \Theta(1)$ for $a \neq a^*$. We first present the results for estimation.

THEOREM EC.1. *If* ***EXP3EG*** *runs with $\alpha \in [0,1]$ and $\delta < 2/e$ in the fully adversarial case, for any $i,j \in [K]$ and $i \neq j$, $\mathbb{E}[\hat{\Delta}_n^{(i,j)}] = \Delta^{(i,j)}$. Moreover, with probability at least $1-\delta$,*

$$|\hat{\Delta}_n^{(i,j)} - \Delta^{(i,j)}| \le \frac{4(2+2K^2(1+e^2))\log\frac{2}{\delta}}{\sqrt{n^{1-\alpha}}}.$$

Particularly, after taking $\delta = \frac{1}{2n^2}$, $\max_{i<j\le K} e(n, \hat{\Delta}_n^{(i,j)}) = \mathcal{O}(\frac{1}{\sqrt{n^{1-\alpha}}})$.

This theorem does not rely on the fixed best assumption, and thus can hold even under the setting where the best arm changes over time.

Proof. We first need to redefine the martingale $M_t^{(i,j)}$ in the proof to Theorem 10. Now we define $M_t^{(i,j)} = \hat{R}_t(i) - \hat{R}_t(i) - \sum_{s=1}^t \Delta_t^{(i,j)}$, and can have

$$\mathbb{E}[M_t^{(i,j)} \mid \mathcal{H}_{t-1}] = M_{t-1}^{(i,j)} + \mathbb{E}\left[\frac{R_t}{\pi_t(i)}\mathbb{I}_{a_t=i} - \frac{R_t}{\pi_t(j)}\mathbb{I}_{a_t=j} - \Delta_t^{(i,j)} \mid \mathcal{H}_{t-1}\right] = M_{t-1}^{(i,j)}.$$

Following the same procedure as Eq. (EC.12), we can have $V_t^{(i,j)} \le \frac{(2+2C')^2(t+1)^{1+\alpha}}{(e-2)}\log\frac{2}{\delta}$.

By Bernstein's inequality, we have, with probability $1-\delta$ for all t:

$$\left|M_t^{(i,j)}\right| \le 4(2+2C')\log\frac{2}{\delta}\sqrt{t^{1+\alpha}},$$

which directly implies the desired results. Q.E.D.

Now we are presenting the regret analysis.

THEOREM EC.2. *Under the fully adversarial case, let* ***EXP3EG*** *run with $\alpha \in [0,1]$ and $\delta = \frac{1}{2n^2}$. The regret is $\mathcal{O}(\sum_{a\in[K]/\{a^*\}} \frac{\max_t \Delta_t(a)\log(n)}{\min_t \Delta_t^2(a)} + \max_t \Delta_t(a) n^{1-\alpha})$.*

Proof. Without loss of generality, we assume the first arm is optimal so that $\mu_1 = \mu_a^*$. Then, we can have $\mathcal{R}(n,\pi) \le \sum_{a\in\{2,\cdots,K\}}(\max_{t\in[n]}\Delta_t(a))\mathbb{E}[T_a(n)]$, where $T_a(n)$ is the random variable of the number of the times that arm action a is played. Define $\tau(a) := \max\{t : a \in \mathcal{A}_t\}$. Following the new

definition of $M_t^{(i,j)}$ we proposed in Theorem EC.1, we already have $V_t^{(i,j)} \le \frac{(2+2C')^2(t+1)^{1+\alpha}}{(e-2)} \log \frac{2}{\delta}$. Define a good event $\mathcal{E}_1 = \{\tau(1) = n\}$. The probability of the inverse event of $\mathcal{E}_1$ can be bounded as:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1^c) &= \mathbb{P}(\tau(1) < n) \\
&= \mathbb{P}(\cup_{t=1}^{n-1}\{\tau(1) = t\}) \\
&\le \sum_{t=1}^{n-1}\sum_{a=2}^{K} \mathbb{P}\left(\tau(1) = t, \tau(a) > t \text{ and } \hat{R}_t(a) - \hat{R}_t(1) \ge 2\sqrt{Ct}\right) \\
&\le \sum_{t=1}^{n-1}\sum_{a=2}^{K} \mathbb{P}\left(\tau(1) = t, \tau(a) > t \text{ and } \hat{R}_t(a) - \hat{R}_t(1) + \sum_{s=1}^{t}\Delta_s(a) \ge 2\sqrt{Ct}\right) \\
&= \sum_{t=1}^{n-1}\sum_{a=2}^{K} \mathbb{P}\left(\tau(1) = t, \tau(a) > t \text{ and } M_t^{(a,1)} \ge 2\sqrt{Ct}\right) \\
&\le 2(n-1)(K-1)\delta,
\end{aligned}$$

the second inequality is the reason why we need the fixed best arm assumption since we need $\sum_{s=1}^{t}\Delta_s(a) \ge 0$. Despite of the different definition of $M_t^{(a,1)}$, Eq. (EC.13) can still hold.

Now, we elaborate on bounding the stopping time $\tau(a)$. We consider the following probability,

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_1, \tau(a) \ge \frac{16C}{\min_{s\in[n]}\Delta_s(a)^2} + 1\right) &= \sum_{t=\frac{16C}{\min_s \Delta_s(a)^2}+1}^{n} \mathbb{P}\left(\mathcal{E}_1, \tau(a) = t\right) \\
&= \sum_{t=\frac{16C}{\min_{s\in[n]} \Delta_s(a)^2}+1}^{n} \mathbb{P}\left(\mathcal{E}_1, \tau(a) = t, \hat{R}_{t-1}(1) - \hat{R}_{t-1}(a) \le 2\sqrt{C(t-1)}\right) \\
&= \sum_{t=\frac{16C}{\min_{s\in[n]} \Delta_s(a)^2}+1}^{n} \mathbb{P}\left(\mathcal{E}_1, \tau(a) = t, M_{t-1}^{(1,a)} \le 2\sqrt{C(t-1)} - \sum_{s=1}^{t-1}\Delta_s(a)\right) \\
&\le \sum_{t=\frac{16C}{\min_{s\in[n]} \Delta_s(a)^2}+1}^{n} \mathbb{P}\left(\mathcal{E}_1, \tau(a) = t, M_{t-1}^{(1,a)} \le -2\sqrt{C(t-1)}\right) \\
&\le 2\left(n - \frac{16C}{\min_{s\in[n]}\Delta_s(a)^2}\right)\delta,
\end{aligned}$$

where the first inequality holds since $t - 1 \ge \frac{16C}{\min_{s\in[n]}\Delta_s(a)^2}$ leads to $\Delta_s(a) \ge \sqrt{\frac{16C}{t-1}}$ for all $s \in [n]$.

Therefore, we can bound $\mathbb{E}[T_a(n)]$ as following

$$\begin{aligned}
\mathbb{E}[T_a(n)] &= \mathbb{E}[T_a(n)\mathbb{I}_{\mathcal{E}_1}] + \mathbb{E}[T_a(n)\mathbb{I}_{\mathcal{E}_1^c}] \\
&\le \mathbb{E}[\tau(a)\mathbb{I}_{\mathcal{E}_1}] + \mathbb{E}\left[\left(\sum_{t=\tau(a)+1}^{n}\mathbb{I}_{A_t=a}\right)\mathbb{I}_{\mathcal{E}_1}\right] + n\mathbb{P}(\mathcal{E}_1^c) \\
&\le \left(\frac{16C}{\min_{s\in[n]}\Delta_s(a)^2} + 1\right) + \sum_{t=1}^{n}\frac{1}{Kt^\alpha} + n\mathbb{P}\left(\mathcal{E}_1, \tau(a) \ge \frac{16C}{\min_{s\in[n]}\Delta_s(a)^2} + 1\right) + n\mathbb{P}(\mathcal{E}_1^c) \\
&\le \frac{16C}{\min_{s\in[n]}\Delta_s(a)^2} + \left(\frac{n^{1-\alpha}}{K(1-\alpha)}\right) \wedge \frac{\log(n)}{K} + 1 + 2Kn^2\delta.
\end{aligned}$$

Thus, the total regret can be bounded as $\mathcal{O}(\sum_{a\in[K]/\{a^*\}} \frac{\max_t \Delta_t(a)\log(n)}{\min_t \Delta_t^2(a)} + \max_t \Delta_t(a) n^{1-\alpha})$. Q.E.D.

E. Technical Details for Section 5

E.1. Proof to Theorem 12

The proof follows from the proof to Theorem 4 of Gao et al. (2019). Define the following events: for $[i]\in K$, let A_i be the event that arm i is eliminated before time t_{m_i}, where

$$m_i = \min\left\{ j\in[M] : \text{arm } i \text{ has been pulled at least } \tau_i^* := \frac{48\log(nK)}{\Delta(i)^2} \text{ times before time } t_j \in \mathcal{T} \right\}.$$

Let B be the event that arm i^* is not eliminated throughout the time horizon n. And the final good event E is defined to be $(\cap_{i=1}^K A_i)\cap B$. We first claim that $\mathbb{P}(E)\geq 1-\frac{2}{TK}$. Thus, the regret $\mathcal{R}(n,\pi)$ under E^c is at most $\mathcal{O}(1)$. Now, we are going to upper bound the regret condition on the event E. Let $\mathcal{I}_0\subset[K]$ be the set of arms that are eliminated after the first batch, $\mathcal{I}_1\subset[K]$ be the set of remaining arms which are eliminated before the last batch, and $\mathcal{I}_2 = [K]\backslash(\mathcal{I}_0\cup\mathcal{I}_1))$. We decompose the regret into three components. We also use $n_{i,j}$ to represent the number of times that arm $i\in[K]$ is played in batch $j\in[M]$.

The total regret incurred by arms in $\mathcal{I}_0$ is at most

$$\sum_{i\in\mathcal{I}_0}\Delta(i)(n_{i,1}+n_{i,M}) = (\max_{i\in\mathcal{I}_0}\Delta(i))\sum_{i\in\mathcal{I}_0} n_{i,1}+n_{i,M} = \mathcal{O}(n^{\frac{1}{M}}+n^{1-\alpha}). \tag{EC.14}$$

The total regret incurred by pulling $i\in\mathcal{I}_i$ is at most

$$\begin{aligned}
\sum_{i\in\mathcal{I}_1}\Delta(i)\left(\frac{t_{m_i}}{K-\sigma_i}+n_{i,M}\right) &= \sum_{i\in\mathcal{I}_1}\frac{1}{\Delta(i)}\frac{\Delta(i)^2 t_{m_i}}{K-\sigma_i}+\mathcal{O}(n^{1-\alpha}) \\
&\leq \sum_{i\in\mathcal{I}_1}\frac{c_b n^{\frac{1}{M}}}{\Delta(i)}\frac{\Delta(i)^2(t_{m_i-1}+1)}{K-\sigma_i}+\mathcal{O}(n^{1-\alpha}) \\
&\leq \sum_{i\in\mathcal{I}_1}\frac{c_b}{\Delta(i)}\frac{48K\log(nK)}{K-\sigma_i}+\mathcal{O}(n^{\frac{1}{M}}+n^{1-\alpha}) \\
&\leq \frac{48K^2 c_b\log(nK)}{\min_i\Delta(i)}+\mathcal{O}(n^{\frac{1}{M}}+n^{1-\alpha}) \\
&= \mathcal{O}\left(\frac{\log(nK)}{\min_i\Delta(i)}+n^{\frac{1}{M}}+n^{1-\alpha}\right),
\end{aligned} \tag{EC.15}$$

where σ_i is the number of arms that are eliminated before arm i is eliminated, the first inequality holds due to the fact that $t_{m_i}\leq c_b^{m_i} n^{\frac{m_i}{M}} = c_b n^{\frac{1}{M}} c_b^{m_i-1} n^{\frac{m_i-1}{M}} \leq c_b n^{\frac{1}{M}}(t_{m_i-1}+1)$, the second inequality follows from the definition of m_i.

The total regret incurred by pulling an arm $i\in\mathcal{I}_2$ which is pulled T_i times is at most

$$\sum_{i\in\mathcal{I}_2}\Delta(i)T_i \leq \sum_{i\in\mathcal{I}_2}\frac{\Delta(i)^2 T_i}{\min_i\Delta(i)}$$

$$\begin{aligned}
&\leq \sum_{i\in\mathcal{I}_2} \frac{48K\log(nK)T_i}{\min_i \Delta(i) t_{M-1}} \\
&\leq \sum_{i\in\mathcal{I}_2} \frac{48K\log(nK)T_i}{\min_i \Delta(i)(c_b^{-1} n^{\frac{M-1}{M}} - 1)} \\
&= \frac{48K\log(nK)n}{\min_i \Delta(i)(c_b^{-1} n^{\frac{M-1}{M}} - 1)} \\
&= \mathcal{O}\left(\frac{n^{\frac{1}{M}}}{\min_i \Delta(i)}\right). \qquad \text{(EC.16)}
\end{aligned}$$

Based on Eqs. (EC.14), (EC.15) and (EC.16), the expected regret is $\widetilde{\mathcal{O}}(n^{\frac{1}{M}} \vee n^{1-\alpha})$.

Now, we are going to prove the error of estimation. If $\alpha > 1 - \frac{1}{M}$, every arm is played deterministic $n_{1,1}$ times. By Hoeffding's inequality,

$$\mathbb{P}\left(\left|\hat{R}_n(a) - \mu_a\right| \leq \sqrt{\frac{2}{n_{1,1}} \log\frac{2}{\delta}}\right) \leq \delta.$$

Thus, by taking $\delta = \frac{1}{\sqrt{2n_{1,1}}}$, $\mathbb{E}[|\hat{R}_n(a) - \mu_a|] \leq \sqrt{\frac{\log(2n_{1,1})}{2n_{1,1}}} + \frac{2}{\sqrt{n_{1,1}}}$. Now, we can have

$$\mathbb{E}[|\hat{\Delta}_n^{(i,j)} - \Delta^{(i,j)}|] \leq \mathbb{E}[|\hat{R}_n(i) - \mu_i|] + \mathbb{E}[|\hat{R}_n(j) - \mu_j|] = \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{n_{1,1}}}\right) = \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{n^{\frac{1}{M}}}}\right). \qquad \text{(EC.17)}$$

If $\alpha \leq 1 - \frac{1}{M}$, $n_{i,M}$ is deterministic and the same for all $i \in [K]\backslash\{i_0\}$ and $n_{i_0,M}$ is also deterministic given i_0. Also, $n_{i,M} < n_{i_0,M}$, and thus $n_{i,M}$ becomes the bottleneck of estimation. By Hoeffding's inequality, $\mathbb{E}[|\hat{R}_n(a) - \mu_a|] = \widetilde{\mathcal{O}}(\frac{1}{\sqrt{n_{i,M}}})$ for any $a \in [K]\backslash\{i_0\}$ and $\mathbb{E}[|\hat{R}_n(i_0) - \mu_{i_0}|] = \widetilde{\mathcal{O}}(\frac{1}{\sqrt{n_{i_0,M}}})$. Therefore, for any i, j,

$$\mathbb{E}[|\hat{\Delta}_n^{(i,j)} - \Delta^{(i,j)}|] = \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{n^{1-\alpha}}}\right) \qquad \text{(EC.18)}$$

Putting Eqs. (EC.17) and (EC.18) together, we finish the proof.

What remains to be shown is $\mathbb{P}(E) \geq 1 - \frac{1}{TK}$. For simplicity, we assume the noise is Guassian distributed, since the sub-Gaussian case follows naturally. If the optimal arm is eliminated by arm i at time t, based on our design, before time t both arms are pulled the same number of times $\tilde{n}$, which can be random. For any given i and fixed $\tilde{n}$,

$$\mathbb{P}\left(\mathcal{N}\left(-\Delta(i), \frac{2}{\tau}\right) \geq \sqrt{\frac{12\log(nK)}{\tau}}\right) \leq \mathbb{P}\left(\mathcal{N}\left(0, \frac{2}{\tau}\right) \geq \sqrt{\frac{12\log(nK)}{\tau}}\right) \leq \frac{1}{(TK)^3}.$$

Therefore, by the union bound,

$$\mathbb{P}(B^c) \leq \sum_{i=1}^{K}\sum_{t=1}^{T}\sum_{\tau=1}^{t} \mathbb{P}(\text{the optimal arm is elimianted by arm } i \text{ at time } t \text{ with } \tau \text{ pulls}) \leq \frac{1}{TK}. \qquad \text{(EC.19)}$$

The event $B \cup A_i^c$ indicates that the optimal arm does not eliminate arm i at time t_{m_i}. Thus, both the optimal arm and arm i have been played at least $\tau > \tau_i^*$ times, which by the definition of τ_i^* implies

$$\Delta(i) \geq \sqrt{\frac{48\log(nK)}{\tau}}.$$

Therefore, for any fixed t_{m_i} and τ,

$$\mathbb{P}\left(\mathcal{N}\left(\Delta(i), \frac{2}{\tau}\right) \leq \sqrt{\frac{12\log(nK)}{\tau}}\right) \leq \mathbb{P}\left(\mathcal{N}\left(0, \frac{2}{\tau}\right) \leq -\sqrt{\frac{12\log(nK)}{\tau}}\right) \leq \frac{1}{(TK)^3}.$$

Again, with the union bound,

$$\mathbb{P}(B \cup A_i^c) \leq \sum_{t_{m_i} \in \mathcal{T}} \sum_{1 \leq \tau \leq T} \frac{1}{(TK)^3} \leq \frac{1}{TK^2}. \tag{EC.20}$$

Together with Eqs. (EC.19) and (EC.20), we derive

$$\mathbb{P}(E^c) \leq \mathbb{P}(B^c) + \sum_{k=1}^{K} \mathbb{P}(B \cap A_i^c) \leq \frac{2}{TK}.$$

We finish the proof. Q.E.D.

F. Technical Details for Section 7

F.1. Proof to Theorem 13

From Lemma 1, we can know that

$$\inf_{\hat{\Delta}_n} \max_{\nu \in \mathcal{E}_0} \mathbb{P}_\nu \left(|\hat{\Delta}_n - \Delta_\nu|_2 \geq \frac{l(n,\delta)}{2} \right) \geq \frac{1}{2}\left[1 - \sqrt{\frac{2l(n,\delta)^2}{3\xi}\mathcal{R}_{\nu_1}(n,\pi)}\right]. \tag{EC.21}$$

Since $l(n,\delta)$ is a valid length of the confidence interval given the confidence level δ, the RHS of Eq. (EC.21) should be at least smaller than $1-\delta$. Otherwise, there is no estimator can guarantee a δ coverage with a confidence interval length $l(n,\delta)$. Therefore, we have that there must exist a suitable ν_1 such that

$$l(n,\delta)\sqrt{\mathcal{R}_{\nu_1}(n,\pi)} \geq (2\delta - 1)\sqrt{\frac{3\xi}{2}} = \Omega(1).$$

Finally, we can have

$$\max_{\nu \in \mathcal{E}_0} \left[l(n,\delta)\sqrt{\mathcal{R}_\nu(n,\pi)} \right] = \Omega(1).$$

Since the above equation holds for any π and $\hat{\Delta}_n$, we finish the proof. Q.E.D.

G. Useful Technical Tools

G.1. Bernstein's Inequality

We made use of the following form of the Bernstein's inequality Freedman (1975), whose proof can be seen in Theorem 8 in Seldin et al. (2013).

THEOREM EC.3 **(Bernstein's Inequality)**. *Let $X_1, X_2, \cdots$ be a martingale difference sequence, such that $|X_t| \le \alpha_t$ for a non-decreasing deterministic sequence $\alpha_1, \alpha_2, \cdots$ with probability 1. Let $M_t := \sum_{\tau=1}^{t} X_\tau$ be martingale. Let $\bar{V}_1, \bar{V}_2, \cdots$ be a deterministic upper bounds on the variance $V_t := \sum_{\tau=1}^{t} \mathbb{E}[X_\tau^2 | X_1, \cdots, X_{\tau-1}]$ of the martingale M_t, such that $\bar{V}_t$-s satisfy $\sqrt{\frac{\ln(\frac{2}{\delta})}{(e-2)\bar{V}_t}} \le \frac{1}{\alpha_t}$. Then with probability greater than $1-\delta$ for all t:*

$$|M_t| \le 2\sqrt{(e-2)\bar{V}_t \ln\frac{2}{\delta}}.$$

H.2. Neyman-Pearson Lemma

THEOREM EC.4 **(Neyman-Pearson Lemma)**. *Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two probability measures. Then for any test ψ, it holds*

$$\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0) \ge \int \min(p_0, p_1).$$

Moreover, the equality holds for the Likelihood Ratio test $\psi^\star = \mathbb{I}(p_1 \ge p_0)$.

COROLLARY EC.1.

$$\inf_{\psi}[\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0)] = 1 - TV(\mathbb{P}_0, \mathbb{P}_1).$$

Proof. Denote that $\mathbb{P}_0$ and $\mathbb{P}_1$ are defined on the probability space $(\mathcal{X}, \mathcal{A})$. By the definition of the total variation distance, we have

$$\begin{aligned}
\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) &= \sup_{R\in\mathcal{A}} |\mathbb{P}_0(R) - \mathbb{P}_1(R)| \\
&= \sup_{R\in\mathcal{A}} \left| \int_R p_0 - p_1 \right| \\
&= \frac{1}{2}\int |p_0 - p_1| \\
&= 1 - \int \min(p_0, p_1) \\
&= 1 - \inf_{\psi}[\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0)],
\end{aligned}$$

where the last equality applies the Neyman-Pearson Lemma, and the fourth equality holds due to the fact that

$$\int |p_0 - p_1| = \int_{p_1 \ge p_0} p_1 - p_0 + \int_{p_1 < p_0} p_0 - p_1$$

$$
\begin{aligned}
&= \int_{p_1 \geq p_0} p_1 + \int_{p_1 < p_0} p_0 - \int \min(p_0, p_1) \\
&= 1 - \int_{p_1 < p_0} p_1 + 1 - \int_{p_1 \geq p_0} p_0 - \int \min(p_0, p_1) \\
&= 2 - 2 \int \min(p_0, p_1).
\end{aligned}
$$

We finish the proof. Q.E.D.

H.3. Pinsker's inequality

THEOREM EC.5 **(Pinsker's inequality)**. *Let* $\mathbb{P}_1$ *and* $\mathbb{P}_2$ *be two probability measures such that* $\mathbb{P}_1 \ll \mathbb{P}_2$. *Then,*

$$TV(\mathbb{P}_1, \mathbb{P}_2) \leq \sqrt{\frac{1}{2} KL(\mathbb{P}_1, \mathbb{P}_2)}.$$