Covert Best Arm Identification of Stochastic Bandits

Meng-Che Chang and Matthieu R. Bloch
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332
Email: {mchang301,matthieu}@gatech.edu

***Abstract*—We study the covert best arm identification problem in which an agent tries to identify the best arm while escaping detection from an adversary. Specifically, the agent should identify the best arm of the bandit with accuracy higher than a predefined requirement as soon as possible and, simultaneously, the adversary's observations induced by pulling effective arms should remain indistinguishable from the observations obtained when no effective arm is pulled. Our main result is the characterization of the exponent γ, which captures the asymptotic exponential decrease of the confidence level with the square-root of the averaged stopping time.**

I. Introduction

In bandit problems, a player pulls an arm on a bandit machine at each time instant and obtains a corresponding reward. A standard objective for the player is to minimize his regret [1], defined as the difference between his rewards and those of the best arm pull strategy, over a fixed time horizon or identify the best arm as soon as possible subject to a certain accuracy constraint [2]. In the regret minimization problem, the player then faces a trade-off between exploiting the most profitable arm identified from past rewards or exploring new arms. Standard algorithms, such as Upper Confidence Bound (UCB) [3] or Active Arm Elimination (AAE) [4], achieve optimal performance by only devoting a small fraction of the time to exploration. On the other hand, the track-and-stop algorithm has been shown to be asymptotically optimal in the best arm identification problem [5].

Different approaches have been proposed to analyze the performance of bandit games in the presence of adversaries. [6], [7] investigate the regret minimization problem in the stochastic bandit setting with bounded but unknown number of corruption on rewards, while [8] studies the best arm identification problem in this setting. [9] explores the case of adversarial bandit, in which rewards are fully decided by an adversary, using the EXP3 algorithm. Different from all of works above, we investigate the situation in which the agent (Alice) would like to identify the best arm of the bandit while keeping the fact that someone is pulling arms unknown to the adversary (Willie).

The problem formulation of this work is inspired by the problem of "low probability of detection," in communication systems studied in [10], [11]. If n denotes the time duration, the ratio of non-innocent symbols transmitted need to be at least $\frac{1}{\sqrt{n}}$ so that the outputs received have a noticeable difference from the outputs when no communication happens. The objective of the agent in this work is to identify the best arm as soon as possible and make the actions of pulling arms covert from Willie's perspective. Similar problem formulations can be found in [12], [13], where the covertness in hypothesis testing problems is analyzed.

This work was supported in part by NSF grant 1955401.

II. Notation

A multi-arm bandit ν with K effective arms and a null arm is a set of distributions $\{\nu_0, ..., \nu_K\}$. We denote ν_k the distribution of the kth arm of the bandit ν. We denote μ the operator that maps a bandit into a vector of means, i.e., $\mu(\nu) = [\mu_0(\nu), ..., \mu_K(\nu)]$, where $\mu_k(\nu)$ is the mean of the distribution ν_k for any $k \in [0; K]$. For any two different bandits ν, ν', $d(\nu, \nu') = ||\mu(\nu) - \mu(\nu')||_\infty$ is the infinity norm of $\mu(\nu) - \mu(\nu')$. For any $k \in [0; K]$ and $t \in \mathbb{N}$, $T_k(t)$ is the number of pulls on the arm k before time $t+1$. For any bandit ν, $i^*(\nu) = \text{argmax}_k\, \mu_k(\nu)$ is the best arm. For any alphabet $\mathcal{U}$, we denote $\mathcal{P}_\mathcal{U}$ the set of all probability distributions on $\mathcal{U}$. For two continuous distributions p, q on some common set $\mathcal{D}$, we define $\chi_2(P||Q) \triangleq \int_{x\in\mathcal{D}} \frac{(p(x)-q(x))^2}{q(x)} p(x)dx$. For any sequence of random variables $\{X_t\}_{t=1}^\infty$, we denote $\mathbf{X}^n \triangleq [X_1, ..., X_n]$ for any $n \in \mathbb{N}$. The notation $r(\delta) = O_{\delta\to 0}(e(\delta))$ means there exists some constant L such that $r(\delta) \leqslant Le(\delta)$ for all δ small enough. Other Landau notations are defined similarly. For any real number x, $|x|^+ \triangleq \max(0, x)$.

III. Problem Formulation

We consider the situation in which an *agent* (Alice) and an *adversary* (Willie) are engaged in the best arm identification problem in the stochastic bandit setting. The distribution of the rewards obtained by Alice and Willie are determined by the *arm* pulled by Alice. Let K be the number of effective arms. We denote $\nu = \{\nu_0, \nu_1, ..., \nu_K\}$ the multi-arm bandit of Alice, and $q = \{q_0, q_1, ...q_K\}$ the multi-arm bandit of Willie. The arm 0 is defined as the null arm. That is, Alice and Willie receive rewards drawn from the distribution ν_0 and q_0, respectively, when no effective arm is pulled. In this work, we assume that the rewards are drawn from Gaussian distributions with unknown means and unit variance, i.e., ν_a and q_a are Gaussian distributions with unit variance for all $a \in \mathcal{A} \triangleq [0; K]$, and we also define $\mathcal{E}_\mathcal{N}$ as the set of all Gaussian bandits with unit variance and K effective arms. In this work, we assume that the information of q is known to the agent but ν is unknown. For any pair of bandits $\nu', \nu'' \in \mathcal{E}_\mathcal{N}$, one cannot distinguish between ν' and ν'' by pulling the null arm, i.e., $D(\nu_0'||\nu_0'') = 0$ for all $\nu', \nu'' \in \mathcal{E}_\mathcal{N}$. Without loss of generality, we assume that the arm 1 is the best arm, i.e., the distribution ν_1 has the largest mean, and ν_0 has zero mean.

The fact that the null arm has zero mean is known by the agent.

The reward distributions of Alice and Willie are connected to a common arm $a \in \mathcal{A}$. Specifically, at each time step t, Alice selects an arm $A_t \in \mathcal{A}$ from the control policy $\pi_t \triangleq P_{A_t|\mathbf{X}^{t-1},\mathbf{A}^{t-1}}$; Alice receives the reward $X_t \sim \nu_{A_k}$; and Willie receives the reward $Z_t \sim q_{A_t}$. Let τ be the stopping time adapted to the filtration $(\mathcal{F}_t)_{=0}^{\infty}$ with $\mathcal{F}_t = \sigma(X_1, A_1, ..., X_t, A_t)$. At time τ, Alice decides on the best arm according to the rule $\psi : \mathbf{X}^\tau \times \mathbf{A}^\tau \mapsto [1; K]$. Let $\delta > 0$ be the parameter defining how accurate the decision on the best arm is. Then, the overall policy $\lambda \triangleq (\{\pi_t\}_{t=1}^{\infty}, \tau, \psi)$ is a triple and has the confidence level δ if

$$\mathbb{P}_\nu(\psi(\mathbf{X}^\tau, \mathbf{A}^\tau) \neq i^*(\nu)) < \delta \text{ for all } \nu \in \mathcal{E}_\mathcal{N}, \tag{1}$$

where $\mathbb{P}_\nu$ is the probability measure under the bandit ν.

The objective of Alice is to identify the best arm as soon as possible while keeping the fact that Alice is pulling arms undetected by (covert from) Willie. The covertness is measured in terms of the divergence. Specifically, we say that the policy λ is η-covert if for all $\nu, q \in \mathcal{E}_\mathcal{N}$,

$$\lim_{\delta\to 0} \mathbb{P}_\nu \left(D(P_{Z^n}||q_0^n) < \eta \text{ for all } 0 \leqslant n \leqslant \tau\right) = 1. \tag{2}$$

We denote by $\Lambda(\eta)$ the set of policies that satisfy the confidence requirement in (1) and are η-covert. We are interested in the exponent $\gamma_{\nu,q}(\lambda)$, which is the ratio between $-\log\delta$ and the square root of the expected stopping time τ, defined as

$$\gamma_{\nu,q}(\lambda) \triangleq \liminf_{\delta\to 0} \frac{-\log\delta}{\sqrt{\mathbb{E}_\nu(\tau)}}, \tag{3}$$

where the expectation $\mathbb{E}_\nu$ is taken over the probability measure $\mathbb{P}_\nu$. Then, we have the following definition of *achievability* in the covert arm identification problem.

Definition 1 (Achievability)**.** *The exponent r is achievable if there exists a policy $\lambda \in \Lambda(\eta)$ such that*

$$\gamma_{\nu,q}(\lambda) \geqslant r. \tag{4}$$

The optimal exponent $\gamma^*_{\nu,q}$ is the supremum of all achievable exponents, i.e.,

$$\gamma^*_{\nu,q} = \sup_{\lambda\in\Lambda(\eta)} \gamma_{\nu,q}(\lambda). \tag{5}$$

The following theorem gives the lower bound on $\gamma^*_{\nu,q}(\eta)$.

Theorem 2. *Let q. For all $\nu \in \mathcal{E}_\mathcal{N}$, if there is no distribution $\bar{P}$ over $\mathcal{A}\setminus\{0\}$ such that $\sum_{a\in\mathcal{A}\setminus\{0\}} \bar{P}(a)q_a = q_0$, then we have*

$$\gamma^*_{\nu,q} \geqslant \sqrt{2\eta} \max_{\bar{P}} \frac{\min_{\nu'\in\mathcal{E}_{Alt}(\nu)} \sum_{a\in\mathcal{A}\setminus\{0\}} \bar{P}(a)D(\nu_a||\nu'_a)}{\sqrt{\chi_2(\sum_{a\in\mathcal{A}\setminus\{0\}} \bar{P}(a)q_a||q_0)}}.$$

We say that the control policy $\{\pi_t\}_{t=1}^{\infty}$ converges almost surely if there exists some $P^* \in \mathcal{P}_\mathcal{A}$ and $C < \infty$ such that for any $a \in \mathcal{A}$ the probability $P_{A_t|\mathbf{X}^{t-1}\mathbf{A}^{t-1}}(a) = P^*(a)$ for all $t \geqslant C|\log\delta|$, and the corresponding overall policy λ is called an almost surely convergent policy. If we restrict ourselves to almost surely convergent policies, the exponent can be upper bounded by the theorem below.

Theorem 3. *Let λ be an almost surely convergent policy, then*

$$\gamma_{\nu,q}(\lambda) \leqslant \sqrt{2\eta} \max_{\bar{P}} \frac{\min_{\nu'\in\mathcal{E}_{Alt}(\nu)} \sum_{a\in\mathcal{A}\setminus\{0\}} \bar{P}(a)D(\nu_a||\nu'_a)}{\sqrt{\chi_2(\sum_{a\in\mathcal{A}\setminus\{0\}} \bar{P}(a)q_a||q_0)}}.$$

The result in Theorem 2 can be interpreted as follows. It is known that the agent can not choose effective arms too often to maintain covertness. Therefore, the policy we propose in Sec. IV is composed of a probability α with which effective arms are chosen and a probability distribution $\bar{P}$ on effective arms. A large value of the divergence term $\min_{\nu'\in\mathcal{E}_{\text{Alt}}(\nu)} \sum_{a\in\mathcal{A}\setminus\{0\}} \bar{P}(a)D(\nu_a||\nu'_a)$ results in a shorter stopping time. On the other hand, $\chi_2(\sum_{a\in\mathcal{A}\setminus\{0\}} \bar{P}(u)q_a||q_0)$ is the chi-square distance between the null distribution and the reward distribution induced by the distribution $\bar{P}$. A smaller chi-square distance implies that Willie is harder to distinguish the reward distribution $\sum_{a\in\mathcal{A}\setminus\{0\}} \bar{P}(a)q_a$ from the null one. This means the agent can choose effective arms more often, i.e. larger α , when the probability of choosing each effective arms is given by $\alpha\bar{P}$. Therefore, the best distribution on effective arms needs to have a good trade-off between maximizing the divergence and minimizing the chi-square distance. Finally, when the covertness constraint η is more relaxed, the agent can choose effective arms more often and speed up the estimation of the best arm.

IV. Proof of Theorem 2

We first specify the policy $\lambda = (\{\pi_t\}_{t=1}^{\infty}, \tau, \psi)$. Let $\hat{\nu}(t)$ be the estimated bandit of ν at the time t, i.e., $\hat{\nu}_k(t) \sim \mathcal{N}\left(\frac{1}{T_k(t)}\sum_{\ell=1}^{t} X_t 1(A_t = k), 1\right)$ for all $k \in \mathcal{A}\setminus\{0\}$, and $\mathcal{E}_{\text{Alt}}(\hat{\nu}(t)) = \{\nu' \in \mathcal{E}_\mathcal{N} : i^*(\hat{\nu}(t)) \cap i^*(\nu') = \emptyset\}$. Then, we define

$$L_t \triangleq \inf_{\nu'\in\mathcal{E}_{\text{Alt}}(\hat{\nu}(t))} \sum_{k=1}^{K} T_k(t)D(\hat{\nu}_k(t)||\nu'_k), \tag{6}$$

and the stopping rule τ is

$$\tau = \inf\{t : L_t > \beta_t(\delta)\}, \tag{7}$$

where $\beta_t(\delta) = k\log(t^2 + t) + f^{-1}(\delta)$ and $f(x) = \exp(K - x)(x/K)^K$. For any $\bar{P} \in \mathcal{P}_{\mathcal{A}\setminus\{0\}}$ and $\nu'' \in \mathcal{E}_\mathcal{N}$, we define

$$\xi(\bar{P}, \nu'') \triangleq \frac{\inf_{\nu'\in\mathcal{E}_{\text{Alt}}(\nu'')} \sum_{k\in[1;K]} \bar{P}(k)D(\nu''_k||\nu'_k)}{\sqrt{\chi_2\left(\sum_{k\in[1;K]} \bar{P}(k)q_k||q_0\right)}}$$

and

$$\zeta(\bar{P}, \nu'') \triangleq \frac{2\eta}{\chi_2\left(\sum_{k\in[1;K]} \bar{P}(k)q_k||q_0\right)} \times \frac{\inf_{\nu'\in\mathcal{E}_{\text{Alt}}(\nu'')} \sum_{k\in[1;K]} \bar{P}(k)D(\nu''_k||\nu'_k)}{|\log 4\delta|}.$$

Algorithm 1: Covert Arm Identification Algorithm

Input: δ, η

Initialization: $s \coloneqq 0, t \coloneqq 0$

while $Z_t < \beta_t(\delta)$ **do**

  if $\arg\min_{k\in[1;K]} T_k(t) \leqslant \sqrt{s}$ **then**

    $\bar{P}_t(a) \coloneqq 1(a = \arg\min_{k\in[1;K]} T_k(t))$ $\forall a \in [1;K]$

  else

    $\bar{P}_t \coloneqq \arg\max_{\bar{P}} \xi(\bar{P}, \hat{\nu}(t))$

  $\alpha_t \coloneqq \zeta(\bar{P}_t, \hat{\nu}(t))$

  $P_{A_{t+1}|X^t,A^t}(a) \coloneqq \begin{cases} 1-\alpha_t & \text{if } a = 0 \\ \alpha_t \bar{P}_t(a) & \text{if } a \neq 0 \end{cases}$

  Draw A_{t+1} from the distribution $P_{A_{t+1}|X^t,A^t}$.

  if $A_{t+1} \neq 0$ **then**

    $s \coloneqq s+1$

  $t \coloneqq t+1$

Output: $i^*(\hat{\nu}(t))$

Then, the policy of determining the arms is defined in Algorithm 1. Note that if there are multiple arms having the same minimum number of times being pulled, then $\text{argmin}_{k\in[1;K]} T_k(t)$ would randomly pick one among them. Finally, the identified best arm at the time τ is $\psi(\mathbf{X}^\tau, \mathbf{A}^\tau) = i^*(\hat{\nu}(\tau))$.

A. Confidence Analysis

The confidence analysis follows from the proof of Lemma 33.7 in [5]. We summarize the main idea behind the proof below. The event $i^*(\hat{\nu}(\tau)) \neq 1$ implies $\nu \in \mathcal{E}_{\text{Alt}}(\hat{\nu}(\tau))$. Then,

$$\mathbb{P}\left(i^*(\hat{\nu}(\tau)) \neq 1\right) \leqslant \mathbb{P}\left(\nu \in \mathcal{E}_{\text{Alt}}(\hat{\nu}(\tau))\right) \tag{8}$$

$$\leqslant \mathbb{P}\left(\sum_{k=1}^{K} T_k(\tau) D(\hat{\nu}_k(\tau)||\nu_k) \geqslant \beta_\tau(\delta)\right). \tag{9}$$

In the case of Gaussian bandit, $D(\hat{\nu}_k(\tau)||\nu_k) = \frac{1}{2}\left(\mu_k(\hat{\nu}(\tau)) - \mu_k(\nu)\right)^2$. The following Lemma [5] provides a concentration bound on the value of $\mu_k(\hat{\nu}(\tau))$.

Lemma 4. *Let $\{X_t\}_{t=1}^{\infty}$ be a sequence of Gaussian random variables with mean μ and unit variance. Let $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$, then*

$$\mathbb{P}\left(\exists n \in \mathbb{N} : \frac{n}{2}(\hat{\mu}_n - \mu)^2 \geqslant \log(1/\delta) + \log(n(n+1))\right) \leqslant \delta$$

for any $\delta > 0$.

Lemma 4 upper bounds the probability that

$$T_k(\tau) D(\hat{\nu}_k(\tau)||\nu_k) \geqslant \log(1/\delta) + \log\left(T_k(\tau)(T_k(\tau)+1)\right)$$

by δ for all $k \in [1;K]$ regardless of $T_k(\tau)$. However, the event on the right hand side of (9) is related to the combination of different divergence terms. The lemma below [5] extends the result in Lemma 4.

Lemma 5. *Let $g : \mathbb{N} \mapsto \mathbb{R}$ be increasing, and for each $k \in [1;K]$, let $S_k = \{S_{k1}, S_{k2}, \cdots\}$ be an infinite sequence of random variables such that for all $\delta \in (0,1)$,*

$$\mathbb{P}\left(\exists n \in \mathbb{N} : S_{1n} \geqslant g(n) + \log(1/\delta)\right) \leqslant \delta. \tag{10}$$

Then, provided that the sequences $\{S_k\}_{k=1}^K$ are independent and $x > 0$,

$$\mathbb{P}\left(\exists \mathbf{s} \in \mathbb{N}^K : \sum_{k=1}^{K} S_{ks_k} \geqslant k g\left(\sum_{k=1}^{K} s_k\right) + x\right) \leqslant \left(\frac{x}{K}\right)^K \exp(K-x). \tag{11}$$

Now, let $s_k = T_k(\tau)$, $S_{ks_k} = D(\hat{\nu}_k(\tau)||\nu_k)$ for all $k \in [1;K]$ and $g(n) = \log(n^2+n)$. Note that the estimations of the empirical distributions of different arms are independent, so are the sequences $\{S_k\}_{k=1}^K$. Applying Lemma 5, we have

$$\mathbb{P}\left(\sum_{k=1}^{K} T_k(\tau) D(\hat{\nu}_k(\tau)||\nu_k) \geqslant \beta_\tau(\delta)\right) \leqslant \mathbb{P}\left(\sum_{k=1}^{K} S_{ks_k} \geqslant k\log(n^2+n) + f^{-1}(\delta)\right) \leqslant \delta$$

B. Divergence Analysis

We first define

$$\bar{P}^* \triangleq \underset{\bar{P}\in\mathcal{P}_{\mathcal{A}\setminus\{0\}}}{\text{argmax}}\ \xi(\bar{P}, \nu)$$

as well as

$$\alpha^* \triangleq \zeta(\bar{P}^*, \nu).$$

Let $\mathcal{B}(\nu,\epsilon) \triangleq \{\nu' \in \mathcal{E}_{\mathcal{N}} : d(\nu,\nu') < \epsilon\}$, $\mathcal{B}(\bar{P}^*,\epsilon) \triangleq \{\bar{P} \in \mathcal{P}_K : ||\bar{P} - \bar{P}^*||_\infty < \epsilon\}$ and $\mathcal{B}(\alpha^*,\epsilon) \triangleq \{\alpha \in \mathbb{R} : |\frac{\alpha}{\alpha^*} - 1| \leqslant \epsilon\}$, and let $\tau_\nu \triangleq \sup\{t : d(\hat{\nu}(t),\nu) > \epsilon\}$, $\tau_{\bar{P}}(\epsilon) \triangleq \sup\{t : ||\bar{P}_t - \bar{P}^*||_\infty > \epsilon\}$ and $\tau_\alpha(\epsilon) \triangleq \sup\{t : |\frac{\alpha_t}{\alpha^*} - 1| > \epsilon\}$. The following lemma shows that $\tau_\nu(\epsilon) = O_{\delta\to 0}(|\log\delta|)$, $\tau_{\bar{P}}(\epsilon) = O_{\delta\to 0}(|\log\delta|)$ and $\tau_\alpha(\epsilon) = O_{\delta\to 0}(|\log\delta|)$ with high probability for all $\epsilon > 0$.

Lemma 6. *For all $\epsilon > 0$, it holds that*

$$\tau_\nu(\epsilon) = O_{\delta\to 0}(|\log\delta|)$$
$$\tau_{\bar{P}}(\epsilon) = O_{\delta\to 0}(|\log\delta|)$$
$$\tau_\alpha(\epsilon) = O_{\delta\to 0}(|\log\delta|)$$

with probability greater than $1-\epsilon$.

For all $0 \leqslant n \leqslant \tau$, the divergence $D(P_{Z^n}||(q_0)^{\otimes n})$ can be expressed as

$$D(P_{Z^n}||(q_0)^{\otimes n}) = \sum_{t=1}^{n} \mathbb{E}_{Z^{t-1}}\left(D(P_{Z_t|Z^{t-1}}||q_0)\right). \tag{12}$$

Let $\tau_{\max}(\epsilon) = \max(\tau_{\bar{P}}(\epsilon), \tau_\alpha(\epsilon))$. Then, for any $n \leqslant \tau$ and fixed $\epsilon > 0$, we have

$D(P_{Z^n}||(q_0)^{\otimes n})$

$$
\begin{aligned}
&\leqslant \sum_{t=1}^{\tau_{\max}(\epsilon)} \mathbb{E}_{Z^{t-1}}\left(D(P_{Z_t|Z^{t-1}}||q_0)\right) + |n-\tau_{\max}(\epsilon)|^+ \\
&\times \sup_{\bar{P}\in\mathcal{B}(\bar{P}^*,\epsilon)} \sup_{\alpha\in\mathcal{B}(\alpha^*,\epsilon)} D\left((1-\alpha)q_0 + \alpha\left(\sum_{k\in[1;K]}\bar{P}(k)q_k\right)\middle\|q_0\right) \\
&\leqslant \sum_{t=1}^{\tau_{\max}(\epsilon)} \mathbb{E}_{Z^{t-1}}\left(D(P_{Z_t|Z^{t-1}}||q_0)\right) + |n-\tau_{\max}(\epsilon)|^+ \\
&\times \sup_{\bar{P}\in\mathcal{B}(\bar{P}^*,\epsilon)} \sup_{\alpha\in\mathcal{B}(\alpha^*,\epsilon)} \left(\frac{\alpha^2}{2}\chi_2\left(\sum_{k\in[1;K]}\bar{P}(k)q_k\middle\|q_0\right) + o(\alpha^2)\right) \\
&\leqslant n\left(\frac{(\alpha^*)^2}{2}\left(\chi_2\left(\sum_{k\in[1;K]}\bar{P}^*(k)q_k\middle\|q_0\right) + h(\epsilon) + o_{\delta\to 0}(1)\right)\right) \\
&+ O_{\delta\to 0}\left(\frac{1}{|\log\delta|}\right), \qquad (13)
\end{aligned}
$$

where $h(\epsilon)$ is some function such that $\lim_{\epsilon\to 0} h(\epsilon) = 0$, and (13) follows from the fact that $\max(\tau_{\bar{P}}(\epsilon), \tau_{\alpha}(\epsilon)) = O_{\delta\to 0}(|\log\delta|)$ and $\mathbb{E}_{Z^{t-1}}\left(D(P_{Z_t|Z^{t-1}}||q_0)\right) = O\left(\frac{1}{|\log\delta|^2}\right)$ for all $t \in \mathbb{N}$. Moreover, we can upper bound the stopping time τ by the following lemma.

Lemma 7. *For any $\epsilon' > 0$, it holds that*

$$
\begin{aligned}
&\mathbb{P}\left(\tau \leqslant \frac{|\log\delta|}{\inf_{\nu'\in\mathcal{E}_{Alt}(\nu)}\sum_{k\in[1;K]}\bar{P}^*(k)D(\nu_k||\nu'_k)} \times \frac{1}{\alpha^*}(1+\epsilon')\right) \\
&\qquad \geqslant 1 - e^{-\Omega_{\delta\to 0}(|\log\delta|)\epsilon'}.
\end{aligned}
$$

Then, with probability $1-\epsilon'$,

$$
\begin{aligned}
&D(P_{Z^n}||(q_0)^{\otimes n}) \\
&\leqslant \frac{\alpha^*}{2}\left(\chi_2\left(\sum_{k\in[1;K]}\bar{P}^*(k)q_k\middle\|q_0\right) + h(\epsilon) + o_{\delta\to 0}(1)\right) \\
&\times \frac{|\log\delta|(1+\epsilon')}{\inf_{\nu'\in\mathcal{E}_{\text{Alt}}(\nu)}\sum_{k\in[1;K]}\bar{P}^*(k)D(\nu_k||\nu'_k)} + O_{\delta\to 0}\left(\frac{1}{|\log\delta|}\right) \\
&\leqslant \eta(1+\epsilon')\left(1 + \frac{h(\epsilon)}{\chi_2\left(\sum_{k\in[1;K]}\bar{P}^*(k)q_k\middle\|q_0\right)} + o_{\delta\to 0}(1)\right) \\
&+ O_{\delta\to 0}\left(\frac{1}{|\log\delta|}\right)
\end{aligned}
$$

for all $n \leqslant \tau$. Since we can make ϵ, ϵ' arbitrarily small, this implies

$$
\lim_{\delta\to 0}\mathbb{P}\left(D(P_{Z^n}||(q_0)^{\otimes n}) \leqslant \eta\right) = 1
$$

for all $n \leqslant \tau$.

C. Characterization of Exponent

Note that Lemma 7 implies that

$$
\mathbb{E}_\nu[\tau] \leqslant \frac{|\log\delta|}{\inf_{\nu'\in\mathcal{E}_{\text{Alt}}(\nu)}\sum_{k\in[1;K]}\bar{P}^*(k)D(\nu_k||\nu'_k)} \times \frac{1}{\alpha^*}(1+\epsilon') \qquad (14)
$$

for any $\epsilon' > 0$ when δ is small enough, and Theorem 2 is obtained by plugging in (14) into the expression of $\gamma_{\nu,q}(\lambda)$.

V. Proof of Theorem 3

Let the control policy π_t converge almost surely to $P^\#$, then there exists some finite constant C such that $\pi_t = P^\#$ for all $t > C|\log\delta|$. We further define

$$
\begin{aligned}
\alpha^\# &\triangleq 1 - P^\#(0) \\
\bar{P}^\# &\triangleq \frac{P^\#}{\alpha^\#}.
\end{aligned}
$$

For any $\nu' \in \mathcal{E}_{\text{Alt}}(\nu)$, we define the event $E \triangleq \{\psi(X^\tau, A^\tau) \notin i^*(\nu')\}$. Then,

$$
\begin{aligned}
2\delta &\geqslant \mathbb{P}_\nu\left(\psi(X^\tau, A^\tau) \notin i^*(\nu)\right) + \mathbb{P}_{\nu'}\left(\psi(X^\tau, A^\tau) \notin i^*(\nu')\right) \\
&\geqslant \mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E) \qquad (15) \\
&\geqslant \frac{1}{2}\exp\left(-\sum_{k=1}^K \mathbb{E}_\nu[T_k(\tau)]D(\nu_k||\nu'_k)\right), \qquad (16)
\end{aligned}
$$

where (16) is from the Bretagnolle–Huber inequality and the divergence decomposition lemma. From [5], we already know that the averaged stopping time $\mathbb{E}_\nu(\tau)$ is $\Omega_{\delta\to 0}(|\log\delta|)$ for all policies in order to satisfy the confidence level constraint when there is no covertness constraint. In our context, one can show that $\mathbb{E}_\nu(\tau) = \Omega_{\delta\to 0}(|\log\delta|^2)$ by the fact that only a small fraction of arms pulled are effective. This implies that for δ small enough,

$$
\begin{aligned}
T_k(\tau) &= \sum_{t=1}^{\tau} 1(A_t = k) \\
&= \sum_{t=1}^{C|\log\delta|} 1(A_t = k) + \sum_{t=C|\log\delta|+1}^{\tau} 1(A_t = k).
\end{aligned}
$$

Then,

$$
\begin{aligned}
\mathbb{E}_\nu[T_k(\tau)] &\leqslant C|\log\delta| + \mathbb{E}_\nu[\tau]P^\#(k) \\
&\leqslant \mathbb{E}_\nu[\tau]P^\#(k)\left(1 + o_{\delta\to 0}(1)\right). \qquad (17)
\end{aligned}
$$

Combining (16) and (17), we obtain

$$
\log\left(\frac{1}{4\delta}\right) \leqslant \mathbb{E}_\nu[\tau]\sum_{k=1}^K P^\#(k)D(\nu_k||\nu'_k)\left(1 + o_{\delta\to 0}(1)\right).
$$

Since the above inequality is true for all $\nu' \in \mathcal{E}_{\text{Alt}}(\nu)$, we can take the infimum of all $\nu' \in \mathcal{E}_{\text{Alt}}(\nu)$ and obtain

$$
\mathbb{E}_\nu[\tau] \geqslant \frac{\log\left(\frac{1}{4\delta}\right)}{\inf_{\nu'\in\mathcal{E}_{\text{Alt}}(\nu)}\sum_{k=1}^K P^\#(k)D(\nu_k||\nu'_k)}\left(1 - o_{\delta\to 0}(1)\right).
$$

This means that there is a positive $\epsilon''' > 0$ such that

$$
\begin{aligned}
&\mathbb{P}\left(\tau \geqslant \frac{\log\left(\frac{1}{4\delta}\right)}{\inf_{\nu'\in\mathcal{E}_{\text{Alt}}(\nu)}\sum_{k=1}^K P^\#(k)D(\nu_k||\nu'_k)}\left(1 - o_{\delta\to 0}(1)\right)\right) \\
&\qquad \geqslant \epsilon'''. \qquad (18)
\end{aligned}
$$

Next, from [11], the covertness constraint implies that

$$
\eta \geqslant D(P_{Z^n}||q_0^{\otimes n}) \geqslant nD(\bar{P}_Z||q_0) \qquad (19)
$$

for all $n \leqslant \tau$, where

$$\bar{P}_Z(z) \triangleq \frac{1}{n}\sum_{t=1}^{n} P_{Z_t}(z) = \frac{1}{n}\sum_{t=1}^{n}\sum_{k=0}^{K}\pi_t(k)q_k(z).$$

Note that

$$\frac{1}{n}\sum_{t=1}^{n}\pi_t(k) = P^{\#}(k)(1+o_{\delta\to 0}(1)) \tag{20}$$

when $n = \Omega_{\delta\to 0}(|\log\delta|^2)$ by the assumption that the policy converges. Then, the covertness constraint and (18) imply that

$$\begin{aligned}\eta &\geqslant \frac{\log\left(\frac{1}{4\delta}\right)}{\inf_{\nu'\in\mathcal{E}_{\text{Alt}}(\nu)}\sum_{k=1}^{K}P^{\#}(k)D(\nu_k||\nu'_k)} D\left(\sum_{k=0}^{K}P^{\#}(k)q_k\middle\|q_0\right)\\ &\geqslant \frac{\log\left(\frac{1}{4\delta}\right)}{\alpha^{\#}\inf_{\nu'\in\mathcal{E}_{\text{Alt}}(\nu)}\sum_{k=1}^{K}\bar{P}^{\#}(k)D(\nu_k||\nu'_k)}\\ &\times\left(\frac{(\alpha^{\#})^2}{2}\chi_2\left(\sum_{k=1}^{K}\bar{P}^{\#}(k)q_k\middle\|q_0\right)-o_{\delta\to 0}((\alpha^{\#})^2)\right)\end{aligned}$$

when δ is sufficiently small. Therefore,

$$\begin{aligned}\alpha^{\#} &\leqslant \frac{2\eta}{\chi_2\left(\sum_{k=1}^{K}\bar{P}\#(k)q_k\middle\|q_0\right)}\\ &\times\frac{\inf_{\nu'\in\mathcal{E}_{\text{Alt}}(\nu)}\sum_{k=1}^{K}\bar{P}^{\#}(k)D(\nu_k||\nu'_k)}{\log(\frac{1}{4\delta})},\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_\nu[\tau] &\geqslant \left(\frac{\log\left(\frac{1}{4\delta}\right)}{\inf_{\nu'\in\mathcal{E}_{\text{Alt}}(\nu)}\sum_{k=1}^{K}\bar{P}^{\#}(k)D(\nu_k||\nu'_k)}\right)^2\\ &\times\frac{\chi_2\left(\sum_{k=1}^{K}\bar{P}^{\#}(k)q_k\middle\|q_0\right)}{2\eta}(1-o_{\delta\to 0}(1))\end{aligned}$$

Taking the infimum over all possible $\bar{P}^{\#}$, we conclude that

$$\gamma^*(\eta) \leqslant \sqrt{2\eta}\sup_{\bar{P}^{\#}}\frac{\inf_{\nu'\in\mathcal{E}_{\text{Alt}}(\nu)}\sum_{k=1}^{K}\bar{P}^{\#}(k)D(\nu_k||\nu'_k)}{\sqrt{\chi_2\left(\sum_{k=1}^{K}\bar{P}^{\#}(k)q_k\middle\|q_0\right)}}.$$

Appendix A
Proof of Lemma 6

We first introduce some notations. Let $N(t) \triangleq \sum_{\ell=1}^{t}1(A_t\neq 0)$ be the number of pulls on effective arms up to time t, and $\tilde{\nu}(s)$ is the estimated bandit with s number of pulls on effective arms. Note that if $N(t)=s$ for some $t\in\mathbb{N}$ and $s\leqslant t$, then $\hat{\nu}(t)=\tilde{\nu}(s)$ because pulling on null arms does not make any difference in estimating the bandit. For any $k\in\mathcal{A}\setminus\{0\}$ and $s\in\mathbb{N}$, we also define $\tilde{T}_k(s)$ as the number of pulls on the arm k when effective arms have been pulled s times. Then, for any $\epsilon>0$, we define $\tilde{\tau_\nu}(\epsilon)$ as

$$\sup\{s : ||\mu(\tilde{\nu}(s))-\mu(\nu)||_\infty \geqslant \epsilon\}, \tag{21}$$

i.e., $\tilde{\tau}_\nu(\epsilon)$ is the greatest value of s such that $||\mu(\tilde{\nu}(s))-\mu(\nu)||_\infty \geqslant \epsilon$. We then show that $\tilde{\tau}_\nu(\epsilon)$ is bounded for all $\epsilon>0$. Define the random variable

$$L \triangleq \inf\left\{\ell\geqslant 1 : d(\mu(\tilde{\nu}(s),\nu)\leqslant\sqrt{\frac{2\log(2\ell Ks(s+1))}{\min_{k\in[1;K]}\tilde{T}_k(s)}}\text{ for all } s\right\}.$$

Then, for any $x\in\mathbb{R}$,

$$\begin{aligned}&\mathbb{P}(L>x)\\ &\leqslant\mathbb{P}\left(d(\mu(\tilde{\nu}(s)),\nu)>\sqrt{\frac{2\log(2xKs(s+1))}{\min_{k\in[1;K]}\tilde{T}_k(s)}}\quad\text{for some } s\right)\\ &\leqslant\sum_{k=1}^{K}\sum_{s=1}^{\infty}\mathbb{P}\left(|\mu_k(\tilde{\nu}(s))-\mu_k(\nu)|>\sqrt{\frac{2\log(2xKs(s+1))}{\tilde{T}_k(s)}}\right)\\ &\leqslant\sum_{k=1}^{K}\sum_{s=1}^{\infty}2\exp(-\log(2xKs(s+1)))\\ &=\frac{1}{x},\end{aligned}$$

which implies

$$\begin{aligned}\mathbb{E}\left((\log L)^2\right) &\leqslant\int_0^\infty\mathbb{P}((\log L)^2>x)dx\\ &=\int_0^\infty\mathbb{P}(L>e^{x/2})dx\\ &\leqslant 2\end{aligned}$$

Therefore, $\mathbb{E}\left((\log L)^2\right)$ is bounded. We can upper bound $\tilde{\tau}_\nu(\epsilon)$ by

$$\tilde{\tau}_\nu(\epsilon)\leqslant 1+\sup\left\{s:\sqrt{\frac{2\log(2LKs(s+1))}{\min_{k\in[1;K]}\tilde{T}_k(s)}}\geqslant\epsilon\right\}.$$

We know that $\tilde{T}_k(s)=\Omega(\sqrt{s})$ for all $k\in[1;K]$ by the construction of Algorithm 1. Then, above upper bound on $\tilde{\tau}_\nu(\epsilon)$ implies that $\mathbb{E}(\tilde{\tau}_\nu(\epsilon))=O(\mathbb{E}((\log L)^2))=O_{\delta\to 0}(1)$ for all $\epsilon>0$. Therefore, $\tilde{\tau}_\nu(\epsilon)<\infty$ is bounded for all $\epsilon>0$. Note that $\alpha_t=\Theta_{\delta\to 0}\left(\frac{1}{|\log\delta|}\right)$ for all t. Then, by concentration inequalities and the fact that $\tilde{\nu}(\epsilon)$ is bounded, this implies that $\tau_\nu(\epsilon)=O_{\delta\to 0}(|\log\delta|)$ with probability $1-\epsilon'$ for any $\epsilon'>0$.

Finally, note that $\bar{P}_t$ is a continuous function of $\hat{\nu}(t)$. For any $\nu''\in\mathcal{E}_{\mathcal{N}}$, we define

$$\bar{P}^*(\nu'')\triangleq\operatorname*{argmax}_{\bar{P}\in\mathcal{P}_K}\frac{\inf_{\nu'\in\mathcal{E}_{\text{Alt}}(\nu'')}\sum_{k\in[1;K]}\bar{P}(k)D(\nu''_k||\nu'_k)}{\sqrt{\chi_2\left(\sum_{k\in[1;K]}\bar{P}(k)q_k||q_0\right)}}$$

as well as

$$\omega(\epsilon)\triangleq\sup\{x:||\bar{P}^*(\nu)-\bar{P}^*(\nu'')||_\infty\leqslant\epsilon\ \forall\nu''\in\mathcal{B}(\nu,x)\}.$$

Then, $\tau_{\bar{P}}(\epsilon)\leqslant\tau_\nu(\omega(\epsilon))=O_{\delta\to 0}(|\log\delta|)$ with probability arbitrarily close to one. The conclusion that $\tau_\alpha(\epsilon)=O_{\delta\to 0}(|\log\delta|)$ can be made by a similar argument.

REFERENCES

[1] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, mar 1985.

[2] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and markov decision processes," in *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2002, pp. 255–270.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer, *Machine Learning*, vol. 47, no. 2/3, pp. 235–256, 2002.

[4] E. Even-Dar, S. Mannor, and Y. Mansour, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," *Journal of Machine Learning Research*, vol. 7, no. 39, pp. 1079–1105, 2006. [Online]. Available: http://jmlr.org/papers/v7/evendar06a.html

[5] C. S. Tor Lattimore, *Bandit Algorithms*. CAMBRIDGE, 2020.

[6] T. Lykouris, V. Mirrokni, and R. Paes Leme, "Stochastic bandits robust to adversarial corruptions," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2018. New York, NY, USA: Association for Computing Machinery, 2018, p. 114–122. [Online]. Available: https://doi.org/10.1145/3188745.3188918

[7] A. Gupta, T. Koren, and K. Talwar, "Better algorithms for stochastic bandits with adversarial corruptions," in *Proceedings of the Thirty-Second Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, A. Beygelzimer and D. Hsu, Eds., vol. 99. PMLR, 25–28 Jun 2019, pp. 1562–1578. [Online]. Available: https://proceedings.mlr.press/v99/gupta19a.html

[8] Z. Zhong, W. C. Cheung, and V. Y. F. Tan, "Probabilistic sequential shrinking: A best arm identification algorithm for stochastic bandits with corruptions," *CoRR*, vol. abs/2010.07904, 2020. [Online]. Available: https://arxiv.org/abs/2010.07904

[9] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, jan 2002.

[10] L. Wang, G. W. Wornell, and L. Zheng, "Fundamental limits of communication with low probability of detection," vol. 62, no. 6, pp. 3493–3503, Jun. 2016.

[11] M. R. Bloch, "Covert communication over noisy channels: A resolvability perspective," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2334–2354, 2016.

[12] M. Tahmasbi and M. R. Bloch, "Active covert sensing," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 840–845.

[13] M.-C. Chang and M. R. Bloch, "Covert sequential hypothesis testing," in *2021 IEEE Information Theory Workshop (ITW)*, 2021, pp. 1–6.