

A Related Work

In this section, we review the other classic problem formulations of the MAB problem. Here we provide a detailed discussion of less directly related works

A.1 Best Arm Identification

Traditional best arm identification (BAI) has been studied for decades [Bechhofer \(1958\)](#). In this setting, there are two possible objectives. First, we have the fixed confidence setting, where the agent aims to minimize the sample time (or the number of samples) while ensuring the best arm is identified with probability greater than $1 - \delta$. Second, there is the fixed budget setting, where the agent has a fixed number of samples and aims to minimize the probability of error.

A.1.1 Fixed Confidence

In [Kaufmann et al. \(2016a\)](#), they introduce a non-asymptotic lower bound for BAI. Subsequently, they propose the Track and Stop algorithm (TAS) that achieves this lower bound asymptotically. However, this algorithm requires the computation of the optimal “proportion” of arm pulls, w^* , at each iteration and requires forced exploration at the order of $\sqrt{\tau}$. The TAS algorithm has since been extended to a variety of other settings [Jourdan et al. \(2023\)](#); [Garivier & Kaufmann \(2021\)](#); [Kato & Ariu \(2021\)](#). In an attempt to ease computation time, [Ménard \(2019\)](#) implements a gradient ascent algorithm to get around having to compute the optimal proportions at each iteration.

In [Degenne et al. \(2019a\)](#), they introduce an optimal algorithm in the non-asymptotic regime. This algorithm formulates the lower bound as a game between two zero-regret algorithms and pulls arms based on the outputted proportions of one player. However, the algorithm is even more computationally expensive than vanilla TAS. For sub-gaussian rewards, [Barrier et al. \(2022\)](#) proposes an optimal algorithm with non-asymptotic optimal upper bound. Here, they also address forced exploration by instead computing the optimal proportions with high probability upper bounds on the empirical means at each iteration. However, the results do not hold for general exponential families.

Prior to these “model-based” methods, there were many “confidence-based” algorithms [Kalyanakrishnan et al. \(2012\)](#); [Kaufmann & Kalyanakrishnan \(2013\)](#); [Gabillon et al. \(2012\)](#) focused on constructing high-probability confidence intervals. The stopping time is then when one confidence interval was disjoint from and greater than all the rest. While these algorithms are more computationally efficient than the aforementioned “model-based” algorithms, they do not perform nearly as well. These algorithms ideas were heavily inspired by the regret minimization setting and the heralded UCB algorithm [Lai & Robbins \(1985\)](#); [Cappé et al. \(2013\)](#); [Lai \(1987\)](#).

A.1.2 Fixed Budget

The budgeted MAB formulation was introduced in [Badanidiyuru et al. \(2013\)](#) as a knapsack problem and allows the reformulation of the MAB problem to an assortment of constraints. As alluded to in [Sinha et al. \(2020\)](#), this framework is not generalizable to problem-dependent cost constraints as needed in our formulation.

From this, successive elimination algorithms emerged [Audibert et al. \(2010\)](#). In these algorithms, the agent pulls every arm at each round then sees if any arms can be eliminated with high probability. The agent then stops when only one arm is remaining. This approach was proven optimal in [Carpentier & Locatelli \(2016\)](#) for compactly supported distributions and provides inspiration for our low-complexity algorithm.

Cost has been introduced into the fixed budget case as a natural extension of the budget. Specifically, non-deterministic cost has been explored in the fixed budget setting for both discrete [Ding et al. \(2013\)](#) and continuous cost [Xia et al. \(2016\)](#). In both cases, the density functions of the costs are compactly supported. Fixed budget BAI can be thought of as the dual problem to our problem

setting. Instead of trying to maximize the confidence subject to a hard cost constraint, we try to minimize the cost subject to a hard confidence constraint.

A.1.3 Best Arm Identification with Safety Constraints

A somewhat similar formulation to the one in this work is the addition of safety constraints to the BAI problem. In Wang et al. (2022), they propose a similar clinical trial example. However, each drug is associated with a dosage and the dosage has an associated safety level. They then attempt to identify the best drug and dosage level for some fixed confidence and safety level. On a similar note, Hou et al. (2022) aims at identifying the best arm subject to a constraint on the variance of the best arm. In some way, our formulation can be viewed as BAI with soft constraints, where the agent is discouraged from certain actions but not forbidden. Additionally, in our formulation we allow for the best arm to be high cost unlike in Hou et al. (2022), where they require that the chosen arm satisfy the variance constraint.

A.2 Regret Minimization

Regret minimization is concerned with minimizing the number of pulls of suboptimal arms or equivalently maximizing the number of times the arm with the highest reward is pulled. Confidence interval based algorithms have proven effective for this setting Lai & Robbins (1985); Lai (1987). This led to fruitful work in finding tighter confidence intervals utilizing Kullback-Liebler divergence Cappé et al. (2013).

A.2.1 Explore Then Commit

Related to BAI, there is a category of regret minimization algorithms that follow an explore-then-commit approach Garivier et al. (2016). In these algorithms, the agent first attempts to identify the best arm as in BAI, then the agent commits to only pulling the identified best arm. However, it was shown in Degenne et al. (2019b) that any ETC algorithm that first utilizes an optimal BAI algorithm will not achieve the optimal regret bounds from Lai & Robbins (1985); Burnetas & Katehakis (1996). A further refined trade-off between best arm identification and regret minimization is studied in Zhang & Ying (2024).

A.2.2 Regret Minimization With Cost

The addition of cost has recently been explored for regret minimization. In Sinha et al. (2020), they introduce the cost as a separate objective, showing that no algorithm can optimally reduce the cost regret and reward regret. Similar to our work, they introduce cost-adapted versions of traditional regret-minimization algorithms. In Yekkehkhany et al. (2021), they introduce cost into the exploration-phase of an ETC algorithm. Here, the agent incurs a cost for each exploration step and is then tasked with minimizing a linear combination of the cost and regret. This is to circumvent the suboptimality of optimal BAI algorithms in regret minimization shown in Degenne et al. (2019b) by introducing regret into the objective.

B Exponential Family

Each of the Bernoulli, Poisson, and Gaussian distributions can be realized as part of the natural exponential family by considering the following choice of parameters:

$$\begin{aligned}
 \text{Bernoulli : } \theta_\mu &= \log\left(\frac{\mu}{1-\mu}\right), & b(\theta_\mu) &= \log(1 + e^{\theta_\mu}), & h(x) &= 1 \\
 \text{Poisson : } \theta_\mu &= \log(\mu), & b(\theta_\mu) &= e^{\theta_\mu}, & h(x) &= \frac{1}{x!}e^{-x} \\
 \text{Gaussian : } \theta_\mu &= \frac{\mu}{\sigma^2}, & b(\theta_\mu) &= \frac{\sigma^2\theta_\mu^2}{2}, & h(x) &= \frac{1}{\sqrt{2\pi}}e^{-x^2/2\sigma^2}
 \end{aligned}$$

C Proofs for the Lower Bound Theorem and Related Corollaries

C.1 Proof of the General Lower Bound Theorem. 1

This section presents a formal proof of Theorem 1. Recall that our goal is to lower bound the expected cumulative cost $\mathbb{E}[J(\tau_\delta)]$, where τ_δ is a stopping time depending on both the randomness of cost samples and the best arm identification algorithm. The following cost decomposition lemma is useful which is an analog to the classic regret decomposition lemma (Lattimore & Szepesvári, 2020, Lemma 4.5).

Lemma 1 (Cost Decomposition Lemma). *For any K -armed stochastic bandit environment, for any algorithm with stopping time τ , the cumulative cost incurred by the algorithm satisfies:*

$$\mathbb{E}[J(\tau)] = \sum_{a=1}^K c_a \cdot \mathbb{E}[N_a(\tau)].$$

The proof of Lemma 1 closely follows the proof of the regret decomposition lemma. The only difference is that now we have to deal with a stopping time τ instead of a fixed time horizon, which can be achieved through applying the tower property carefully. With the help of the cost decomposition lemma, we are ready to prove Theorem 1.

Proof of Theorem 1. Let $\delta \in (0, 1)$ be the confidence level, let $\mu \in \mathcal{S}$ be any bandit instance that we study, and consider a δ -PAC algorithm. For any time index $t \geq 1$, denote by $N_a(t)$ the (random) number of draws of arm a up to time t .

Similar to the proof of Theorem 1 in Garivier & Kaufmann (2016), we first invoke the 'transportation' lemma (Kaufmann et al., 2016a, Lemma 1), (Garivier & Kaufmann, 2016, Theorem 1) to build a relationship between the expected number of draws $N_a(t)$ and the Kullback-Leibler divergence from bandit instance μ to another bandit model λ which has a different best arm: for any $\lambda \in \mathcal{S}$ such that $a^*(\lambda) \neq a^*(\mu)$, we have:

$$\sum_{a=1}^K d(\mu_a, \lambda_a) \mathbb{E}_\mu[N_a(\tau_\delta)] \geq \sup_{\mathcal{E} \in \mathcal{F}_{\tau_\delta}} d(\mathbb{P}_\mu(\mathcal{E}), \mathbb{P}_\lambda(\mathcal{E})).$$

Selecting the event $\mathcal{E} = \{\hat{a} = 1\}$, we notice that our considered algorithm is δ -PAC, so we require $\mathbb{P}_\mu(\mathcal{E}) \geq 1 - \delta$ and $\mathbb{P}_\lambda(\mathcal{E}) \leq 1 - \delta$, since the optimal arm for instance μ is arm 1 but the optimal arm for instance λ is not arm 1. Together with the symmetry and monotonicity of the d function, we have:

$$d(\delta, 1 - \delta) \leq \sum_{a=1}^K d(\mu_a, \lambda_a) \mathbb{E}_\mu[N_a(\tau_\delta)].$$

Then, we can choose the instance $\lambda \in \text{Alt}(\mu)$ which has a different best arm other than instance μ , to achieve the inf of the right hand side. With a little manipulation, we have:

$$d(\delta, 1 - \delta) \leq \inf_{\lambda \in \text{Alt}(\mu)} \sum_{a=1}^K d(\mu_a, \lambda_a) \mathbb{E}_\mu[N_a(\tau_\delta)] = \inf_{\lambda \in \text{Alt}(\mu)} \mathbb{E}_\mu[J(\tau_\delta)] \left(\sum_{a=1}^K \frac{\mathbb{E}_\mu[N_a(\tau_\delta)]}{\mathbb{E}_\mu[J(\tau_\delta)]} d(\mu_a, \lambda_a) \right).$$

Now we use the cost decomposition lemma 1 to decompose the cumulative cost in the denominator, and then we have:

$$\begin{aligned} d(\delta, 1 - \delta) &= \mathbb{E}_\mu[J(\tau_\delta)] \inf_{\lambda \in \text{Alt}(\mu)} \left(\sum_{a=1}^K \frac{c_a \mathbb{E}_\mu[N_a(\tau_\delta)]}{c_a \mathbb{E}_\mu[J(\tau_\delta)]} d(\mu_a, \lambda_a) \right) \\ &= \mathbb{E}_\mu[J(\tau_\delta)] \inf_{\lambda \in \text{Alt}(\mu)} \left(\sum_{a=1}^K \frac{1}{c_a} \frac{c_a \mathbb{E}_\mu[N_a(\tau_\delta)]}{\sum_{a=1}^K c_a \mathbb{E}_\mu[N_a(\tau_\delta)]} d(\mu_a, \lambda_a) \right). \end{aligned}$$

We let $w_a = \frac{c_a \mathbb{E}_\mu [N_a(\tau_\delta)]}{\sum_{a=1}^K c_a \mathbb{E}_\mu [N_a(\tau_\delta)]}$ be the proportion of cost incurred by arm a which sums up to 1. Then, we can further upper bound the above expression by taking the sup over \mathbf{w} as:

$$d(\delta, 1 - \delta) \leq \mathbb{E}_\mu [J(\tau_\delta)] \sup_{\mathbf{w} \in \Sigma_K} \inf_{\text{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^K \frac{w_a}{c_a} d(\mu_a, \lambda_a) \right) \leq \mathbb{E}_\mu [J(\tau_\delta)] T^*(\boldsymbol{\mu})^{-1}.$$

Thus, we can finally derive the lower bound on the cumulative cost as desired:

$$\mathbb{E}_\mu [J(\tau_\delta)] \geq T^*(\boldsymbol{\mu}) d(\delta, 1 - \delta).$$

□

C.2 Proof of Lemma 1

Proof of Lemma. 1. For any stopping time τ , we apply the tower property to condition on the stopping time τ as follows:

$$\mathbb{E}[J(\tau)] = \mathbb{E} \left[\sum_{t=1}^{\tau} C_t \right] = \mathbb{E} \left[\sum_{a=1}^K \sum_{t=1}^{\tau} C_t \mathbf{1}_{A_t=a} \right] = \sum_{a=1}^K \mathbb{E} \left[\mathbb{E} \left[\sum_{t=1}^{\tau} C_t \mathbf{1}_{A_t=a} \middle| \tau \right] \right].$$

On event $\mathbf{1}_{A_t=a}$, C_t is a random variable with expectation c_a and independent of the stopping time τ , so we have:

$$\mathbb{E} \left[\sum_{t=1}^{\tau} C_t \mathbf{1}_{A_t=a} \middle| \tau \right] = c_a \mathbb{E} \left[\sum_{t=1}^{\tau} \mathbf{1}_{A_t=a} \middle| \tau \right] = c_a \mathbb{E} [N_a(\tau) | \tau].$$

Therefore, we can put everything back and reuse the tower property as:

$$\sum_{a=1}^K \mathbb{E} \left[\mathbb{E} \left[\sum_{t=1}^{\tau} C_t \mathbf{1}_{A_t=a} \middle| \tau \right] \right] = \sum_{a=1}^K c_a \mathbb{E} [\mathbb{E} [N_a(\tau) | \tau]] = \sum_{a=1}^K c_a \mathbb{E} [N_a(\tau)].$$

□

C.3 Jensen-Shannon Divergence Form of Lower Bound

Before proving the lower bound theorems in special Gaussian bandit models, we first provide an alternative characterization of the lower bound together with the instance-dependent constants $T^*(\boldsymbol{\mu})$ and $\mathbf{w}^*(\boldsymbol{\mu})$. Similar to [Garivier & Kaufmann \(2016\)](#), for two expected rewards μ_1 and μ_2 , we introduce the identically parameterized Jensen-Shannon divergence $I_\alpha(\mu_1, \mu_2)$ as follows:

$$I_\alpha(\mu_1, \mu_2) := \alpha d(\mu_1, \alpha\mu_1 + (1 - \alpha)\mu_2) + (1 - \alpha) d(\mu_2, \alpha\mu_1 + (1 - \alpha)\mu_2). \quad (2)$$

The following Proposition characterizes the instance-dependent minimax problem from Theorem. 1 using this Jensen-Shannon divergence:

Proposition 3. For every $w \in \Sigma_K$,

$$\inf_{\lambda \in \text{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^K \frac{w_a}{c_a} d(\mu_a, \lambda_a) \right) = \min_{a \neq 1} \left(\frac{w_1}{c_1} + \frac{w_a}{c_a} \right) I_{\frac{w_1/c_1}{w_1/c_1 + w_a/c_a}}(\mu_1, \mu_a).$$

It follows that

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_K} \min_{a \neq 1} \left(\frac{w_1}{c_1} + \frac{w_a}{c_a} \right) I_{\frac{w_1/c_1}{w_1/c_1 + w_a/c_a}}(\mu_1, \mu_a),$$

$$w^*(\boldsymbol{\mu}) = \operatorname{argmax}_{w \in \Sigma_K} \min_{a \neq 1} \left(\frac{w_1}{c_1} + \frac{w_a}{c_a} \right) I_{\frac{w_1/c_1}{w_1/c_1 + w_a/c_a}}(\mu_1, \mu_a).$$

Proof of Proposition. 3. Let μ such that $\mu_1 > \mu_2 \geq \dots \geq \mu_K$. Using the fact that

$$\text{Alt}(\mu) = \bigcup_{a \neq 1} \{\lambda \in \mathcal{S} : \lambda_a > \lambda_1\},$$

one has

$$\begin{aligned} T^*(\mu)^{-1} &= \sup_{w \in \Sigma_K} \min_{a \neq 1} \inf_{\lambda \in \mathcal{S} : \lambda_a > \lambda_1} \sum_{a=1}^K \frac{w_a}{c_a} d(\mu_a, \lambda_a) \\ &= \sup_{w \in \Sigma_K} \min_{a \neq 1} \inf_{\lambda \in \mathcal{S} : \lambda_a \geq \lambda_1} \left[\frac{w_1}{c_1} d(\mu_1, \lambda_1) + \frac{w_a}{c_a} d(\mu_a, \lambda_a) \right], \end{aligned}$$

where the second equality is true because the inner inf is achieved when $\lambda_{a'} = \mu_{a'}$ for all $a' \notin \{1, a\}$, and thus $d(\lambda_{a'}, \mu_{a'}) = 0$. Let $f(\lambda_1, \lambda_a)$ to be the function inside the bracket as follows:

$$f(\lambda_1, \lambda_a) = \frac{w_1}{c_1} d(\mu_1, \lambda_1) + \frac{w_a}{c_a} d(\mu_a, \lambda_a).$$

Then optimizing $f(\lambda_1, \lambda_a)$ under the constraint $\lambda_a \geq \lambda_1$ is a convex optimization problem that can be solved analytically. The minimum is obtained for

$$\lambda_1 = \lambda_a = \frac{w_1/c_1}{w_1/c_1 + w_a/c_a} \mu_1 + \frac{w_a/c_a}{w_1/c_1 + w_a/c_a} \mu_a$$

and value of $T^*(\mu)^{-1}$ can be rewritten $\left(\frac{w_1}{c_1} + \frac{w_a}{c_a}\right) I_{\frac{w_1/c_1}{w_1/c_1 + w_a/c_a}}(\mu_1, \mu_a)$, using the function I_α defined in (2). \square

C.4 Proof of Corollary. 1

Proof for Corollary. 1. We start with the expression of $T^*(\mu)^{-1}$. Recall its expression in two armed bandit models with $\mu_1 > \mu_2$:

$$T^*(\mu)^{-1} = \sup_{w_1 + w_2 = 1} \inf_{\lambda_2 > \lambda_1} \frac{w_1}{c_1} d(\mu_1, \lambda_1) + \frac{w_2}{c_2} d(\mu_2, \lambda_2).$$

Recall that under unit-variance Gaussian bandit models, the d-divergence of two distributions with mean μ and λ is simply $\frac{1}{2}(\mu - \lambda)^2$, so we have:

$$T^*(\mu)^{-1} = \sup_{w_1 \in [0,1]} \inf_{\lambda_2 > \lambda_1} \frac{w_1 (\mu_1 - \lambda_1)^2}{2c_1} + \frac{(1 - w_1) (\mu_2 - \lambda_2)^2}{2c_2}.$$

Fix w_1 and solve the inner optimization problem, we have:

$$\lambda_1 = \lambda_2 = \frac{c_2 w_1}{c_2 w_1 + c_1 (1 - w_1)} \mu_1 + \frac{c_1 (1 - w_1)}{c_2 w_1 + c_1 (1 - w_1)} \mu_2.$$

Substitute the expression of λ_1 and λ_2 into the outer optimization, we have:

$$\begin{aligned} T^*(\mu)^{-1} &= \sup_{w_1 \in [0,1]} \frac{w_1 (1 - w_1)^2 c_1 (\mu_1 - \mu_2)^2}{2 (c_2 w_1 + c_1 (1 - w_1))^2} + \frac{w_1^2 (1 - w_1) c_2 (\mu_1 - \mu_2)^2}{2 (c_2 w_1 + c_1 (1 - w_1))^2} \\ &= \frac{(\mu_1 - \mu_2)^2}{2} \sup_{w_1 \in [0,1]} \frac{w_1 (1 - w_1)}{c_2 w_1 + c_1 (1 - w_1)}. \end{aligned}$$

Then, solving this maximization problem gives us:

$$w_1 = \frac{\sqrt{c_1}}{\sqrt{c_1} + \sqrt{c_2}}.$$

Substitute the above expression into $T^*(\boldsymbol{\mu})^{-1}$, we can easily get:

$$T^*(\boldsymbol{\mu})^{-1} = \frac{(\mu_1 - \mu_2)^2}{2(\sqrt{c_1} + \sqrt{c_2})^2}.$$

Based on this expression of $T^*(\boldsymbol{\mu})^{-1}$ and Theorem. 1, the lower bound of cumulative cost for Gaussian two-armed bandit model with unit variance is simply:

$$\mathbb{E}[J(\tau_\delta)] \geq T^*(\boldsymbol{\mu}) d(\delta, 1 - \delta) = \frac{2(\sqrt{c_1} + \sqrt{c_2})^2 d(\delta, 1 - \delta)}{(\mu_1 - \mu_2)^2}.$$

□

C.5 Lower Bound on Another Simplified MAB Model

Similar to the 2-armed Gaussian case, we can get an explicit lower bound on a slightly more general MAB model. Namely, we can handle an arbitrary amount of suboptimal arms provided that they share reward and cost distribution.

Corollary 2. *Let $\delta \in (0, 1)$. For any δ -PAC algorithm and any K -armed unit-variance Gaussian bandit model with mean $\mu_1 > \mu_2 = \dots = \mu_K$ and $c_1 \neq c_2 = \dots = c_K$, we have for any $a \in \mathcal{A}$:*

$$\mathbb{E}_{\boldsymbol{\mu} \times c}[J(\tau_\delta)] \geq \frac{2\left(c_1 + c_a + \frac{K\sqrt{c_1 c_a}}{\sqrt{K-1}}\right)}{(\mu_1 - \mu_a)^2} d(\delta, 1 - \delta).$$

Proof of Corollary. 2. Let $\Delta = |\mu_1 - \mu_a|$. By Proposition. 3, and together with the definition of Jensen-Shannon divergence in Eq. (2) for Gaussian bandits, we can write $T^*(\boldsymbol{\mu})$ in closed form as:

$$\begin{aligned} T^*(\boldsymbol{\mu})^{-1} &= \max_{w:w_1+(K-1)w_a=1} \left(\frac{w_1}{c_1} + \frac{w_a}{c_a} \right) \left(\alpha \frac{(1-\alpha)^2(\mu_1 - \mu_a)^2}{2} + (1-\alpha) \frac{\alpha^2(\mu_1 - \mu_a)^2}{2} \right) \\ &= \max_{w:w_1+(K-1)w_a=1} \left(\frac{w_1}{c_1} + \frac{w_a}{c_a} \right) \alpha(1-\alpha) \frac{\Delta^2}{2}, \end{aligned}$$

where $\alpha = \frac{w_1/c_1}{w_1/c_1 + w_a/c_a}$ and $\Delta = \mu_1 - \mu_a$. So further, we have:

$$T^*(\boldsymbol{\mu})^{-1} = \max_{w:w_1+(K-1)w_a=1} \frac{\frac{w_1 w_a}{c_1 c_a} \Delta^2}{\frac{w_1}{c_1} + \frac{w_a}{c_a} 2}$$

substitute w_1 with $1 - (K-1)w_a$ and take derivatives to maximize it, we can get:

$$\begin{aligned} w_1^* &= \frac{\sqrt{c_1}}{\sqrt{c_1} + \sqrt{(K-1)c_a}} = \frac{\sqrt{(K-1)c_1}}{\sqrt{(K-1)c_1} + (K-1)\sqrt{c_a}} \\ w_a^* &= \frac{\sqrt{c_a}}{\sqrt{K-1}(\sqrt{c_1} + \sqrt{c_a(K-1)})} = \frac{\sqrt{c_a}}{\sqrt{(K-1)c_1} + (K-1)\sqrt{c_a}}. \end{aligned}$$

So substitute the expression in, we have:

$$T^*(\boldsymbol{\mu})^{-1} = \frac{\sqrt{K-1}}{\left(\sqrt{(K-1)c_1} + \sqrt{c_a}\right) \left(\sqrt{(K-1)c_a} + \sqrt{c_1}\right)} \frac{\Delta^2}{2} = \frac{\Delta^2}{2\left(c_1 + c_a + \frac{K\sqrt{c_1 c_a}}{\sqrt{K-1}}\right)}.$$

Thus the lower bound is

$$\mathbb{E}\left[\sum_a c_a N_\tau(a)\right] \geq \frac{2d(\delta, 1 - \delta)}{\Delta^2} \left(c_1 + c_a + \frac{K\sqrt{c_1 c_a}}{\sqrt{K-1}}\right).$$

□

D Computing the Optimal Proportion \mathbf{w}^*

D.1 Characterization of Optimal Proportion \mathbf{w}^*

As in [Garivier & Kaufmann \(2016\)](#), for every $a \in \{2, \dots, K\}$, we define the $g_a(x) : \mathbb{R} \rightarrow \mathbb{R}$ function which resembles the expression of instance dependent constant $T^*(\boldsymbol{\mu})$ as follows:

$$g_a(x) = (1+x)I_{\frac{1}{1+x}}(\mu_1, \mu_a).$$

The function g_a is a strictly increasing one-to-one mapping from $[0, +\infty)$ onto $[0, d(\mu_1, \mu_a))$ ([Garivier & Kaufmann, 2016](#), Theorem 5). Next, we define its inverse function $x_a(y) : \mathbb{R} \rightarrow \mathbb{R}$ as follows:

$$x_a(y) = g_a^{-1}(y).$$

So $x_a(y)$ is also a strictly increasing one-to-one mapping from $[0, d(\mu_1, \mu_a))$ to $[0, +\infty)$. Using this notation, we can simplify the notation for $T^*(\boldsymbol{\mu})$ according to Proposition 3 as follows:

$$\begin{aligned} T^*(\boldsymbol{\mu})^{-1} &= \operatorname{argmax}_{w \in \Sigma_K} \min_{a \neq 1} \left(\frac{w_1}{c_1} + \frac{w_a}{c_a} \right) I_{\frac{w_1/c_1}{w_1/c_1 + w_a/c_a}}(\mu_1, \mu_a) \\ &= \operatorname{argmax}_{w \in \Sigma_K} \left[\frac{w_1}{c_1} \min_{a \neq 1} g_a \left(\frac{w_a/c_a}{w_1/c_1} \right) \right] \end{aligned}$$

It is clear that the solution to the max-min problem resides in the critical point where all $g_a(\frac{w_a/c_a}{w_1/c_1})$ are equal to each other. Then, the problem of computing the optimal proportion \mathbf{w}^* becomes solving the balancing point for $g_a(x)$ functions. Let x_1^* being a constant 1. In Lemma 2, we show that this balancing point of the $g_a(x)$ function is exactly the solution $\mathbf{x}^* = (x_1^*, \dots, x_a^*)$ to the following maximization problem:

$$\operatorname{argmax}_{(x_1, \dots, x_K)} \left[\frac{\min_{a \neq 1} g_a(x_a)}{c_1 + c_2 x_2 + \dots + c_K x_K} \right]. \quad (3)$$

Lemma 2. *For any two different sub-optimal arm a and b , we have that $g_a(x_a^*) = g_b(x_b^*)$, which means:*

$$\operatorname{argmax}_{w \in \Sigma_K} \left[\frac{w_1}{c_1} \min_{a \neq 1} g_a \left(\frac{w_a/c_a}{w_1/c_1} \right) \right] = \operatorname{argmax}_{(x_1, \dots, x_K)} \left[\frac{\min_{a \neq 1} g_a(x_a)}{c_1 + c_2 x_2 + \dots + c_K x_K} \right].$$

Let $y^* = g_a(x_a^*)$ for all arm $a \in \mathcal{A}$ and notice that $x_a^* = x_a(y^*)$, so we can rewrite the optimization problem. (3) as follows:

$$y^* = \operatorname{argmax}_y G(y) := \frac{y}{c_1 + c_2 x_2(y) + \dots + c_K x_K(y)}.$$

To solve y^* , one simply can let the derivative of $G(y)$ equal 0. After obtaining y^* , it is simple to obtain \mathbf{w}^* through the inverse mapping $x_a(y)$. Then, we summarize the characterization of \mathbf{w}^* in the following Theorem. 4.

Theorem 4. *The optimal proportion $\mathbf{w}^* = (w_1^*, \dots, w_K^*)$ can be computed through:*

$$w_a^* = \frac{c_a x_a(y^*)}{c_1 + c_2 x_2(y^*) + \dots + c_K x_K(y^*)},$$

where y^* is the unique solution to $F_{\boldsymbol{\mu}, \mathbf{c}}(y) = 1$ for

$$F_{\boldsymbol{\mu}, \mathbf{c}} : y \mapsto \sum_{a=2}^K \frac{c_a \cdot d(\mu_1, \frac{\mu_1 + x_a(y)\mu_a}{1+x_a(y)})}{c_1 \cdot d(\mu_a, \frac{\mu_1 + x_a(y)\mu_a}{1+x_a(y)})}.$$

D.2 Computing w^* Efficiently

Since both $g_a(x)$ and $F_{\mu,c}(y)$ functions are continuous, we can solve the zero points of $F_{\mu,c}(y) - 1$ and $g_a(x) - y^*$ efficiently, e.g., through bisection methods or Newton's iteration. We sum up the algorithm to compute w^* in Algorithm. 3. Here, the function `ZeroPoint(\cdot)` takes the input of a continuous monotonic function and outputs its zero point using an arbitrary method.

Algorithm 3: ComputeProportions(μ, c)

Input: expected reward μ ; expected cost c .

compute the maximum point y^* of function $G(y)$ through its derivative's zero point:

$y^* \leftarrow \text{ZeroPoint}(F_{\mu,c}(y) - 1)$;

for $a \in [K]$ **do**

 compute the solution: $x_a^* \leftarrow \text{ZeroPoint}(g_a(x) - y^*)$

for $a \in [K]$ **do**

 transform back to the optimal proportion: $w_a^* = \frac{c_a x_a(y^*)}{c_1 + c_2 x_2(y^*) + \dots + c_K x_K(y^*)}$;

return optimal proportion $w^* = (w_1^*, w_2^*, \dots, w_K^*)$

D.3 Proofs for Lemma. 2 and Theorem. 4

Proof of Lemma. 2. We now show that all the $g_a(x_a^*)$ have to be equal. Let \mathcal{B} be the set of arms that the optimal $g_b(x_b^*)$ function on the optimal solution x^* is not among the minimum of all arms, i.e.,

$$\mathcal{B} = \left\{ b \in \{2, \dots, K\} : g_b(x_b^*) = \min_{a \neq 1} g_a(x_a^*) \right\}$$

and $\mathcal{A} = \{2, \dots, K\} \setminus \mathcal{B}$ to be its complement which achieves the minimum $g_a(x)$ function on x^* . Assume $\mathcal{A} \neq \emptyset$. For all $a \in \mathcal{A}$ and $b \in \mathcal{B}$, one has $g_a(x_a^*) > g_b(x_b^*)$. Using the continuity of the g functions and the fact that they are strictly increasing, there exists $\epsilon > 0$ such that

$$\forall a \in \mathcal{A}, b \in \mathcal{B}, \quad g_a(x_a^* - \epsilon/c_a|\mathcal{A}|) > g_b(x_b^* + \epsilon/c_b|\mathcal{B}|) > g_b(x_b^*)$$

Introducing $\bar{x}_a = x_a^* - \epsilon/c_a|\mathcal{A}|$ for all $a \in \mathcal{A}$ and $\bar{x}_b = x_b^* + \epsilon/c_b|\mathcal{B}|$ for all $b \in \mathcal{B}$, there exists $b \in \mathcal{B}$:

$$\frac{\min_{a \neq 1} g_a(\bar{x}_a)}{c_1 + c_2 \bar{x}_2 + \dots + c_K \bar{x}_K} = \frac{g_b(x_b^* + \epsilon/c_b|\mathcal{B}|)}{c_1 + c_2 x_2^* + \dots + c_K x_K^*} > \frac{g_b(x_b^*)}{c_1 + c_2 x_2^* + \dots + c_K x_K^*} = \frac{\min_{a \neq 1} g_a(x_a^*)}{c_1 + c_2 x_2^* + \dots + c_K x_K^*},$$

which contradicts the fact that x^* belongs to the solution of the optimization problem (3). Hence $\mathcal{A} = \emptyset$ and there exists $y^* \in [0, d(\mu_1, \mu_2))$ such that

$$\forall a \in \{2, \dots, K\}, g_a(x_a^*) = y^* \Leftrightarrow x_a^* = x_a(y^*).$$

□

Proof of Theorem. 4. Recall the $G(y)$ function as follows:

$$G(y) := \frac{y}{c_1 + c_2 x_2(y) + \dots + c_K x_K(y)}.$$

From (3) and Lemma 2, we know that $y^* = \underset{y}{\operatorname{argmax}} G(y)$. Therefore, we can solve for y^* analytically.

From Garivier & Kaufmann (2016), we can take the derivative of $x_a(y)$ and obtain:

$$x'_a(y) = 1/d(\mu_a, m_a(x_a(y))), \quad \text{where } m_a(x) = \frac{\mu_1 + x\mu_a}{1 + x}$$

Therefore, we can also compute the derivative of $G(y)$ as follows:

$$G'(y) = \frac{(c_1 + c_2 x_2(y) + \dots + c_K x_K(y)) - \sum_{a=2}^K \frac{c_a y}{d(\mu_a, m_a(x_a(y)))}}{(c_1 + c_2 x_2(y) + \dots + c_K x_K(y))^2}.$$

The condition $G'(y) = 0$ gives us that

$$\sum_{a=2}^K \frac{c_a y}{d(\mu_a, m_a(x_a(y)))} = c_1 + c_2 x_2(y) + \dots + c_K x_K(y).$$

Thus, y^* is the solution to the following

$$\sum_{a=2}^K \frac{c_a \cdot d(\mu_1, m_a(x_a(y)))}{c_1 \cdot d(\mu_a, m_a(x_a(y)))} = 1,$$

which is exactly Theorem. 4 suggests. After obtaining y^* , we can recover w_a^* as follows:

$$w_1^* = \frac{c_1}{c_1 + c_2 x_2(y^*) + \dots + c_K x_K(y^*)},$$

$$w_a^* = \frac{c_a x_a(y^*)}{c_1 + c_2 x_2(y^*) + \dots + c_K x_K(y^*)}, \quad \forall a \in \{2, \dots, K\}.$$

□

E Proof Roadmap of Theorem. 2

We first show that the empirical cost proportion converges to the optimal proportion w^* in Section. F and in Theorem. 5. Then, we prove a weaker version of Theorem. 2 which states that the cost performance of CTAS matches the lower bound asymptotically with probability 1, i.e., Theorem. 6 in Sec. G. Finally, we prove Theorem. 2 in Sec. H.

F Asymptotic Convergence of Cost Proportions $\widehat{w}(t)$ for CTAS

Theorem 5. *Following the sampling rule of the CTAS algorithm in Algorithm. 1, we have:*

$$\mathbb{P}_{\mu \times c} \left(\lim_{t \rightarrow \infty} \frac{\widehat{c}_a(t) N_a(t)}{J(t)} = w_a^* \right) = 1.$$

F.1 Proof of Theorem 5

We now provide a complete proof of Theorem. 5 which shows that the empirical cost proportion converges to the optimal proportion w_a^* for all actions almost surely. We follow the notation that $\widehat{w}_a(t) = \widehat{c}_a(t) N_a(t) / J(t)$ for all arm a and let $w^*(\mu, c) = \{w_a^*(\mu, c)\}_{a \in \mathcal{A}}$ to be the output of Algorithm. 3 when the input is (μ, c) . First of all, it is easy to see that $w^*(\mu, c)$ is a continuous function.

Lemma 3. *For every μ and c , $w^*(\mu, c)$ is continuous at (μ, c) .*

Proof of Lemma. 3. Since the cost of each arm is bounded, we inherit the same properties of $F_{\mu, c}(y)$ as $F_{\mu}(y)$ from Garivier & Kaufmann (2016). Namely, $\frac{d}{dy} F_{\mu, c}(y^*) \neq 0$. Since y^* is solved by letting $F_{\mu, c}(y^*) = 1$, it is clear that y^* is continuous in terms of the input (μ, c) . By composition, $x_a(y^*)$ is a continuous of (μ, c) and consequently so is $w^*(\mu, c)$ as desired. □

Let $\widehat{w}^*(t) = \{w_a^*(\widehat{\mu}(t), \widehat{c}(t))\}_{a \in \mathcal{A}}$ for simplicity. The following lemma shows that almost surely, $\widehat{w}^*(t)$ converges to w_a^* as $t \rightarrow \infty$, which is ensured by the forced exploration mechanism of CTAS.

Lemma 4. *The sampling rule of the CTAS algorithm presented in algorithm. 1 ensures $\mathbb{P}(\lim_{t \rightarrow \infty} \hat{\mathbf{w}}^*(t) = \mathbf{w}^*) = 1$.*

Proof of Lemma. 4. We first show that the number of pulls $N_a(t)$ for all arms will increase to infinity as t increases to infinity, when we ignore the stopping rule. Specifically, we first show that for any arm a and any time t , $N_a(t) \geq \sqrt{t} - 1$.

For all positive integer n , we define integer $t_n = \inf\{t \in \mathbb{N} : \sqrt{t} \geq n\}$. We divide the time horizon into frames, and the n -th frame is denoted by $\mathcal{I}_n = \{t_n, \dots, t_{n+1} - 1\}$. When n is large enough, we have $|\mathcal{I}_n| > K$. For all $t \in \mathcal{I}_n$, we have $n \leq \sqrt{t} < n + 1$. According to the proof of Lemma. 17 from Garivier & Kaufmann (2016) and using a similar induction argument, it is easy to show that $\forall t \in \mathcal{I}_n$, $N_a(t) \geq n$. This is because the algorithm will be forced to pull arms which has not been pulled n times in the first K steps in this frame. Therefore, we have for $t \in \mathcal{I}_n$, $N_a(t) \geq n > \sqrt{t} - 1$ for all action a . This indicates that $N_a(t) \geq \sqrt{t} - 1$ for all t and a . Then, we have $N_a(t) \rightarrow \infty$ as $t \rightarrow \infty$.

By the Strong Law of Large Numbers, we have that $\hat{\boldsymbol{\mu}}(t) \rightarrow \boldsymbol{\mu}$ almost surely and $\hat{\mathbf{c}}(t) \rightarrow \mathbf{c}$ almost surely as $N_a(t)$ increases to infinity. Let \mathcal{E} be the event that: $\mathcal{E} = \{\hat{\boldsymbol{\mu}}(t) \rightarrow \boldsymbol{\mu}\} \cap \{\hat{\mathbf{c}}(t) \rightarrow \mathbf{c}\}$, we have $P(\mathcal{E}) = 1$. Then by continuity of $\mathbf{w}^*(\boldsymbol{\mu}, \mathbf{c})$ from Lemma. 3, we have $\mathbf{w}^*(t) \rightarrow \mathbf{w}^*$ for any sample path in \mathcal{E} . This means $\mathbb{P}(\lim_{t \rightarrow \infty} \hat{\mathbf{w}}^*(t) = \mathbf{w}^*) = 1$. \square

Recall that our CTAS algorithm will pull the arm which has the largest cost deficit, so that it could contribute more cost and bring the empirical proportion $\hat{\mathbf{w}}(t)$ closer to the estimate $\hat{\mathbf{w}}^*(t)$. So when the estimate of estimate $\hat{\mathbf{w}}^*(t)$ is close enough to the optimal proportion \mathbf{w}^* , it should be intuitive that the empirical cost proportion $\hat{\mathbf{w}}(t)$ is also close to the optimal proportion due to the negative feedback mechanism. The next Lemma. 5 ensures this property, which is essential in proving Theorem. 5.

Lemma 5. *For every $\epsilon > 0$, if there exists t_0 such that under the sampling rule of the CTAS algorithm, we have:*

$$\max_a |\hat{w}_a^*(t) - w_a^*| < \epsilon, \quad \forall t \geq t_0,$$

then, the CTAS algorithm will ensure that there exists a constant $t_\epsilon \geq t_0$ such that for all $t \geq t_\epsilon$:

$$\left| \frac{\hat{c}_a(t)N_a(t)}{J(t)} - w_a^* \right| \leq 3(K-1)\epsilon.$$

Now we are ready to prove Theorem. 5 with the help of Lemma. 4 and Lemma. 5.

Proof of Theorem 5. Let \mathcal{E} be the set of sample paths such that $\{\lim_{t \rightarrow \infty} \hat{\mathbf{w}}^*(t) = \mathbf{w}^*\}$ holds. According to Lemma. 4, \mathcal{E} holds almost surely and there exists a constant t_0 such that for every $\omega \in \mathcal{E}$ and $t \geq t_0$, we have:

$$\max_a |w_a^*(\hat{\boldsymbol{\mu}}(t)) - w_a^*(\boldsymbol{\mu})| < \frac{\epsilon}{3(K-1)}.$$

Therefore, using Lemma 5, there exists $t_\epsilon > t_0$ such that for every $t \geq t_\epsilon$, we have:

$$\max_a \left| \frac{\hat{c}_a(t)N_a(t)}{J(t)} - w_a^*(\boldsymbol{\mu}) \right| < \epsilon,$$

which implies $\lim_{t \rightarrow \infty} \hat{w}_a(t) \rightarrow w_a^*$ for every $\omega \in \mathcal{E}$. \square

F.2 Proof of Lemma. 5

The proof of Lemma. 5 takes inspiration from the tracking results in. (Garivier & Kaufmann, 2016, Lemma. 8), which is motivated by Antos et al. (2008). However in our case, the total cost $J(t)$ and the empirical cost estimation $\hat{c}(t)$ are both random variables which doesn't appear in the original proofs. It requires more delicate analysis.

Proof of Lemma. 5. Define \mathcal{E} to be the set that $\mathbf{w}^*(t) \rightarrow \mathbf{w}^*$, $\hat{\mu}(t) \rightarrow \mu$, and $\hat{c}(t) \rightarrow c$ hold. And by Lemma. 4 and the strong law of large numbers, \mathcal{E} holds almost surely. Define $E_{a,t} := \hat{c}_a(t)N_a(t) - J(t)w_a^*$ to be the cost overhead compared to the optimal proportion \mathbf{w}^* . It is easy to see that:

$$\sum_{a=1}^K E_{a,t} = \sum_{a=1}^K \hat{c}_a(t)N_a(t) - \sum_{a=1}^K J(t)w_a^* = J(t) - J(t) = 0.$$

For each action a , we have $E_{a,t} \leq \max_a E_{a,t}$ on one hand. On the other hand, we can lower bound its overhead as follows:

$$E_{a,t} = - \sum_{a' \neq a} E_{a',t} \geq -(K-1) \max_a E_{a,t}.$$

Therefore, the maximum overhead in absolute value can be bounded as follows:

$$\max_a |E_{a,t}| \leq (K-1) \max_a E_{a,t}.$$

Then, it suffice to bound $E_{a,t}$. By the boundedness of cost, there exists a constant t_0 such that for $t \geq t_0$, we have all the following hold:

$$\sqrt{t} < 2\epsilon t \leq 2J(t)\epsilon \leq \frac{2J(t)\epsilon}{\hat{c}_a(t)}, \quad \forall a \in \mathcal{A}; \quad \max_a |\hat{w}_a^*(t) - w_a^*| \leq \epsilon; \quad \max_a |\hat{\mu}_a(t) - \mu_a| \leq \epsilon; \quad \max_a |\hat{c}_a(t) - c_a| \leq \epsilon.$$

This means $\hat{c}_a(t)\sqrt{t} \leq 2J(t)\epsilon$ for $t \geq t_0$. If at time t , the algorithm picks $A_{t+1} = a$ to explore, it means either this arm a is under-explored, i.e., $N_a(t) \leq \sqrt{t}$, or the arm has the largest cost deficit, i.e.,

$$a = \operatorname{argmax}_{a'} J(t)\hat{w}_{a'}^*(t) - \hat{c}_{a'}(t)N_{a'}(t).$$

In the first case, we can bound the cost overhead as follows:

$$E_{a,t} = \hat{c}_a(t)N_a(t) - J(t)w_a^* \leq \hat{c}_a(t)\sqrt{t} - J(t)w_a^* \leq \hat{c}_a(t)\sqrt{t} \leq 2J(t)\epsilon.$$

In the second case, we have:

$$\begin{aligned} \hat{c}_a(t)N_a(t) - J(t)\hat{w}_a^*(t) &= \min_{a'} \hat{c}_{a'}(t)N_{a'}(t) - J(t)\hat{w}_{a'}^*(t) = \min_a E_{a,t} + J(t)(\hat{w}_a^*(t) - w_a^*) \\ &\leq \min_a E_{a,t} + J(t)\epsilon \\ &\leq 2J(t)\epsilon, \end{aligned}$$

where the first inequality holds for every sample path on \mathcal{E} , and the second inequality is due to $\min_a E_{a,t} \leq 0$. Therefore, we have $\{A_{t+1} = a\} \subset \{E_{a,t} \leq 2J(t)\epsilon, \forall t \geq t_0\}$. We next show by induction that $E_{a,t} \leq \max\{E_{a,t_0}, 2J(t)\epsilon + 1\}$ for all $t \geq t_0$. For our base case, let $t = t_0$, and the statement clearly holds. So let $t \geq t_0$ such that we assume $E_{a,t} \leq \max\{E_{a,t_0}, 2J(t)\epsilon + 1\}$, then if $E_{a,t} \leq 2J(t)\epsilon$, we have:

$$\begin{aligned} E_{a,t+1} &= E_{a,t} + C(t+1)\mathbf{1}_{\{A_{t+1}=a\}} - C(t+1)w_a^* \leq E_{a,t} + C(t+1)\mathbf{1}_{\{E_{a,t} \leq 2J(t)\epsilon\}} - C(t+1)w_a^* \\ &= E_{a,t} + C(t+1)(1 - w_a^*) \leq 2J(t)\epsilon + 1 \leq 2J(t+1)\epsilon + 1 \leq \max\{E_{a,t_0}, 2J(t+1)\epsilon + 1\}, \end{aligned}$$

where the first inequality is due to $\{A_{t+1} = a\} \subset \{E_{a,t} \leq 2J(t)\epsilon, \forall t \geq t_0\}$, and the second inequality uses the induction assumption. If $E_{a,t} > 2J(t)\epsilon$, the indicator is zero and we have:

$$E_{a,t+1} \leq E_{a,t} - C(t+1)w_a^* \leq \max\{E_{a,t_0}, 2J(t)\epsilon + 1\} - C(t+1)w_a^* \leq \max\{E_{a,t_0}, 2J(t+1)\epsilon + 1\},$$

which concludes the induction that $E_{a,t} \leq \max\{E_{a,t_0}, 2J(t)\epsilon + 1\}$ for all $t \geq t_0$. Substitute the definition of $E_{a,t}$ in and we will have $\widehat{c}_a(t)N_a(t) - J(t)w_a^* \leq \max\{E_{a,t_0}, 2J(t)\epsilon + 1\}$, which indicates:

$$\frac{\widehat{c}_a(t)N_a(t)}{J(t)} - w_a^* \leq \max\left(\frac{E_{a,t_0}}{J(t)}, 2\epsilon + \frac{1}{J(t)}\right) \leq \max\left(\frac{t_0}{J(t)}, 2\epsilon + \frac{1}{J(t)}\right).$$

Since $J(t) \rightarrow \infty$ as t increases, there exists a constant $t_\epsilon \geq t_0$ such that for any $t \geq t_\epsilon$, we have:

$$\frac{\widehat{c}_a(t)N_a(t)}{J(t)} - w_a^* \leq \max\left(\frac{t_0}{J(t)}, 2\epsilon + \frac{1}{J(t)}\right) \leq 3\epsilon.$$

Therefore, for every $t \geq t_\epsilon$, we have:

$$\max_a \left| \frac{\widehat{c}_a(t)N_a(t)}{C(t)} - w_a^* \right| = \max_a \frac{|E_{a,t}|}{J(t)} \leq (K-1) \max_a \frac{E_{a,t}}{J(t)} = (K-1) \max_a \left(\frac{\widehat{c}_a(t)N_a(t)}{J(t)} - w_a^* \right) \leq 3(K-1)\epsilon.$$

□

G Asymptotic Cumulative Cost Upper Bound for CTAS

Theorem 6 (Almost Sure Cost Upper Bound). *Let $\delta \in [0, 1]$ and $\alpha \in [1, e/2]$. Using the Chernoff's stopping rule with $\beta(t, \delta) = \log(\mathcal{O}(t^\alpha)/\delta)$, the CTAS algorithm ensures:*

$$\mathbb{P}_{\mu \times c} \left(\limsup_{\delta \rightarrow 0} \frac{J(\tau_\delta)}{\log(1/\delta)} \leq \alpha T^*(\mu) \right) = 1.$$

In this section, we give a proof of Theorem. 6 and Theorem. 2 which characterizes the upper bound of cumulative cost asymptotically. Before proving the Theorems, we present the following technical lemma which is useful in these proofs and can be checked easily.

Lemma 6. *For every $\alpha \in [1, e/2]$, for any two constants $c_1, c_2 > 0$,*

$$x = \frac{\alpha}{c_1} \left[\log \left(\frac{c_2 e}{c_1^\alpha} \right) + \log \log \left(\frac{c_2}{c_1^\alpha} \right) \right]$$

is such that $c_1 x \geq \log(c_2 x^\alpha)$.

G.1 Proof of Theorem. 6

In this section, we first prove the almost sure upper bound from Theorem. 6 with the help of Theorem. 5.

Proof of Theorem. 6. Let \mathcal{E} be the event that all concentrations regarding reward, cost, and empirical proportion holds, i.e.,

$$\mathcal{E} = \left\{ \forall a \in \mathcal{A}, \frac{c_a N_a(t)}{J(t)} \xrightarrow[t \rightarrow \infty]{} w_a^* \right\} \cap \left\{ \widehat{\mu}(t) \xrightarrow[t \rightarrow \infty]{} \mu \right\} \cap \left\{ \widehat{c}(t) \xrightarrow[t \rightarrow \infty]{} c \right\}$$

By Theorem. 5 and the law of large numbers, \mathcal{E} is of probability 1. On \mathcal{E} , there exists t_0 such that for all $t \geq t_0$, $\widehat{\mu}_1(t) > \max_{a \neq 1} \widehat{\mu}_a(t)$ due to the concentration of empirical reward, and thus the Chernoff stopping statistics can be re-written as:

$$\begin{aligned} Z(t) &= \min_{a \neq 1} Z_{1,a}(t) = \min_{a \neq 1} N_1(t) d(\widehat{\mu}_1(t), \widehat{\mu}_{1,a}(t)) + N_a(t) d(\widehat{\mu}_a(t), \widehat{\mu}_{1,a}(t)) \\ &= J(t) \cdot \left[\min_{a \neq 1} \left(\frac{N_1(t)}{J(t)} + \frac{N_a(t)}{J(t)} \right) I_{\frac{N_1(t)/J(t)}{N_1(t)/J(t) + N_a(t)/J(t)}}(\widehat{\mu}_1(t), \widehat{\mu}_a(t)) \right]. \end{aligned}$$

By Lemma 3, for all $a \geq 2$, the mapping $(\mathbf{w}, \boldsymbol{\lambda}, \mathbf{c}) \rightarrow \left(\frac{w_1}{c_1} + \frac{w_a}{c_a}\right) I_{\frac{w_1/c_1}{(w_1/c_1 + w_a/c_a)}}(\lambda_1, \lambda_a)$ is continuous at $(w^*(\boldsymbol{\mu}), \boldsymbol{\mu}, \mathbf{c})$. Therefore, for all $\epsilon > 0$ there exists $t_1 \geq t_0$ such that for all $t \geq t_1$ and all $a \in \{2, \dots, K\}$,

$$\left(\frac{N_1(t)}{C(t)} + \frac{N_a(t)}{C(t)}\right) I_{\frac{N_1(t)/C(t)}{N_1(t)/C(t) + N_a(t)/C(t)}}(\hat{\mu}_1(t), \hat{\mu}_a(t)) \geq \frac{w_1^*/c_1 + w_a^*/c_a}{1 + \epsilon} I_{\frac{w_1^*/c_1}{w_1^*/c_1 + w_a^*/c_a}}(\mu_1, \mu_a)$$

Hence, for any $t \geq t_1$, we have:

$$Z(t) \geq J(t) \min_{a \neq 1} \frac{w_1^*/c_1 + w_a^*/c_a}{1 + \epsilon} I_{\frac{w_1^*/c_1}{w_1^*/c_1 + w_a^*/c_a}}(\mu_1, \mu_a) = \frac{J(t)}{(1 + \epsilon)T^*(\boldsymbol{\mu})}.$$

Consequently, we can bound the cumulative cost a stopping time τ_δ as follows:

$$\begin{aligned} J(\tau_\delta) &= J\left(\inf\{t \in \mathbb{N} : Z(t) \geq \beta(t, \delta)\}\right) \leq J(t_1) \vee J\left(\inf\{t \in \mathbb{N} : J(t)(1 + \epsilon)^{-1}T^*(\boldsymbol{\mu})^{-1} \geq \log(Bt^\alpha/\delta)\}\right) \\ &\leq J(t_1) \vee (1 + \epsilon)T^*(\boldsymbol{\mu}) \left(\log\left(\frac{B}{\ell^\alpha \delta}\right) + \alpha \log(J(\tau_\delta))\right) + \mathcal{O}(1), \end{aligned}$$

for some positive constant B from Proposition 1. Using Lemma 6, it follows that on \mathcal{E} , for $\alpha \in [1, e/2]$

$$J(\tau_\delta) \leq J(t_1) \vee \alpha(1 + \epsilon)T^*(\boldsymbol{\mu}) \left[\log\left(\frac{Be((1 + \epsilon)T^*(\boldsymbol{\mu}))^\alpha}{\ell^\alpha \cdot \delta}\right) + \log\log\left(\frac{B((1 + \epsilon)T^*(\boldsymbol{\mu}))^\alpha}{\ell^\alpha \cdot \delta}\right)\right] + \mathcal{O}(1).$$

Thus we have:

$$\limsup_{\delta \rightarrow 0} \frac{J(\tau_\delta)}{\log(1/\delta)} \leq (1 + \epsilon)\alpha T^*(\boldsymbol{\mu}).$$

Letting ϵ go to zero concludes the proof. \square

G.2 Asymptotic Expectation Optimality

In this section, we provide the proof of Theorem 2 which characterizes the expected cumulative cost upper bound for CTAS.

Proof of Theorem 2. Without loss of generality, we assume that for every $a \in \{1, \dots, K\}$, $c_a \in [\ell, 1]$ with $\ell > 0$. To ease the notation, we assume that the bandit model $\boldsymbol{\mu}$ is such that $\mu_1 > \mu_2 \geq \dots \geq \mu_K$. Let $\epsilon > 0$. From the continuity of w^* in $\boldsymbol{\mu}$, there exists two continuous functions $\alpha(\epsilon)$ and $\beta(\epsilon)$ with $\lim_{\epsilon \rightarrow 0} \alpha(\epsilon) = 0$ and $\lim_{\epsilon \rightarrow 0} \beta(\epsilon) = 0$, and we have $\alpha(\epsilon) \leq (\mu_1 - \mu_2)/4$ such that

$$\mathcal{I}_\epsilon := [\mu_1 - \alpha(\epsilon), \mu_1 + \alpha(\epsilon)] \times \dots \times [\mu_K - \alpha(\epsilon), \mu_K + \alpha(\epsilon)],$$

and $\beta(\epsilon)$ small enough so that

$$\mathcal{J}_\epsilon := [c_1 - \beta(\epsilon), c_1 + \beta(\epsilon)] \times \dots \times [c_K - \beta(\epsilon), c_K + \beta(\epsilon)]$$

is such that for all $(\boldsymbol{\mu}', \mathbf{c}') \in \mathcal{I}_\epsilon \times \mathcal{J}_\epsilon$,

$$\max_a |w_a^*(\boldsymbol{\mu}', \mathbf{c}') - w_a^*(\boldsymbol{\mu}, \mathbf{c})| \leq \epsilon$$

In particular, whenever $\hat{\boldsymbol{\mu}}(t) \in \mathcal{I}_\epsilon$, the empirical best arm is $\hat{a}_t = 1$. Let $T \in \mathbb{N}$ and define $h(T) := T^{1/4}$ and the event

$$\mathcal{E}_T(\epsilon) = \bigcap_{t=h(T)}^T \{\hat{\boldsymbol{\mu}}(t) \in \mathcal{I}_\epsilon\} \cap \{\hat{\mathbf{c}}(t) \in \mathcal{J}_\epsilon\}$$

We first present a lemma showing that the event $\mathcal{E}_T(\epsilon)$ is a high probability event as follows:

Lemma 7. *There exists constants B and C such that*

$$\mathbb{P}(\mathcal{E}_T^c) \leq BT \exp(-CT^{1/8}).$$

The proof of this lemma will be delayed. We now proceed in proving Theorem 2. By Lemma 5, we have that there exists some T_ϵ such that for $T \geq T_\epsilon$, it holds on \mathcal{E}_T that

$$\forall t \geq \sqrt{T}, \max_a \left| \frac{\widehat{c}_a(t)N_a(t)}{J(t)} - w_a^*(\boldsymbol{\mu}, \mathbf{c}) \right| \leq 3(K-1)\epsilon$$

On the event \mathcal{E}_T , it holds for $t \geq h(T)$ that $\widehat{a}_t = 1$ and the Chernoff stopping statistic rewrites

$$\begin{aligned} \max_a \min_{b \neq a} Z_{a,b}(t) &= \min_{a \neq 1} Z_{1,a}(t) = \min_{a \neq 1} N_1(t)d(\widehat{\mu}_1(t), \widehat{\mu}_{1,a}(t)) + N_a(t)d(\widehat{\mu}_a(t), \mu_{1,a}(t)) \\ &= J(t) \cdot \left[\min_{a \neq 1} \left(\frac{N_1(t)}{J(t)} + \frac{N_a(t)}{J(t)} \right) I_{\frac{N_1(t)/J(t)}{N_1(t)/J(t) + N_a(t)/J(t)}}(\widehat{\mu}_1(t), \widehat{\mu}_a(t)) \right] \\ &= J(t) \cdot g\left(\widehat{\boldsymbol{\mu}}(t), \left(\frac{\widehat{c}_a(t)N_a(t)}{J(t)} \right)_{a=1}^K\right), \end{aligned}$$

where we introduce the function

$$g(\boldsymbol{\mu}', \mathbf{c}', \mathbf{w}') = \min_{a \neq 1} \left(\frac{w'_1}{c'_1} + \frac{w'_a}{c'_a} \right) I_{\frac{w'_1/c'_1}{w'_1/c'_1 + w'_a/c'_a}}(\mu'_1, \mu'_a)$$

Using Lemma 5, for $T \geq T_\epsilon$, introducing

$$C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c}) = \inf_{\substack{\boldsymbol{\mu}': \|\boldsymbol{\mu}' - \boldsymbol{\mu}\| \leq \alpha(\epsilon) \\ \mathbf{c}': \|\mathbf{c}' - \mathbf{c}\| \leq \beta(\epsilon) \\ \mathbf{w}': \|\mathbf{w}' - \mathbf{w}^*(\boldsymbol{\mu}, \mathbf{c})\| \leq 3(K-1)\epsilon}} g(\boldsymbol{\mu}', \mathbf{c}', \mathbf{w}'),$$

where $\alpha(\epsilon)$ and $\beta(\epsilon)$ are two continuous functions such that $\lim_{\epsilon \rightarrow 0} \alpha(\epsilon) = 0$ and $\lim_{\epsilon \rightarrow 0} \beta(\epsilon) = 0$. On the event \mathcal{E}_T , it holds that for every $t \geq \sqrt{T}$, we have:

$$\max_a \min_{b \neq a} Z_{a,b}(t) \geq J(t) \cdot C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c}).$$

Let $T \geq T_\epsilon$. On \mathcal{E}_T , we have:

$$\begin{aligned} \min(J(\tau_\delta), J(T)) &\leq J(\sqrt{T}) + \sum_{t=\sqrt{T}}^T C_t \cdot \mathbf{1}_{(\tau_\delta > t)} \leq J(\sqrt{T}) + \sum_{t=\sqrt{T}}^T C_t \cdot \mathbf{1}_{(\max_a \min_{b \neq a} Z_{a,b}(t) \leq \beta(T, \delta))} \\ &\leq J(\sqrt{T}) + \sum_{t=\sqrt{T}}^T C_t \cdot \mathbf{1}_{(J(t) \cdot C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c}) \leq \beta(T, \delta))} \leq J(\sqrt{T}) + \frac{\beta(T, \delta)}{C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})}. \end{aligned}$$

For every sample path ω , consider $T_0(\delta, \omega) := \inf \left\{ T \mid J(\sqrt{T}) + \frac{\beta(T, \delta)}{C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})} \leq J(T, \omega) \right\}$. Then for every $T \geq \max(T_0(\delta, \omega), T_\epsilon)$, that is for every T such that $J(T, \omega) \geq \max\{J(T_0(\delta, \omega)), T_\epsilon\}$, we have that $\min\{J(\tau_\delta, \omega), J(T, \omega)\} < J(T, \omega)$, and thus $J(\tau_\delta, \omega) < J(T, \omega)$. So it means $J(T_0(\delta, \omega)) + T_\epsilon$ is an upper bound of $J(\tau_\delta, \omega)$ and next we intend to bound $J(T_0(\delta, \omega))$ for every $\omega \in \mathcal{E}_T$. Consider

$$F(\eta, \omega) = \inf \left\{ T \in \mathbb{N} \mid J(T, \omega) - J(\sqrt{T}) \geq J(T, \omega)/(1 + \eta) \right\} = \inf \left\{ T \in \mathbb{N} \mid \frac{\eta}{1 + \eta} J(T, \omega) \geq J(\sqrt{T}) \right\}.$$

Then it is easy to see that

$$F(\eta, \omega) \leq \inf \left\{ T \in \mathbb{N} \mid \frac{\eta \ell T}{1 + \eta} \geq \sqrt{T} \right\} =: F'(\eta).$$

Therefore, we have by the boundedness of cost:

$$\begin{aligned} J(T_0(\delta, \omega)) &\leq J(F(\eta)) + J\left(\inf\left\{T \mid \frac{1}{C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})} \log\left(\frac{BT^\alpha}{\delta}\right) \leq \frac{J(T, \omega)}{1+\eta}\right\}\right) \\ &\leq J(F'(\eta)) + J\left(\inf\left\{T \mid \frac{J(T, \omega)C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})}{1+\eta} \geq \log\left(\frac{BT^\alpha}{\delta}\right)\right\}\right) \end{aligned}$$

where B is some constant from Proposition 1. Define

$$T_1(\delta, \omega) := \inf\left\{T \mid \frac{J(T, \omega)C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})}{1+\eta} \geq \log\left(\frac{BT^\alpha}{\delta}\right)\right\}$$

By Lemma 6, we have:

$$\begin{aligned} J(T_1(\delta, \omega)) &\leq \frac{1+\eta}{C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})} \log\left(\frac{BT_1(\delta, \omega)^\alpha}{\delta}\right) + \mathcal{O}(1) \leq \frac{1+\eta}{C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})} \log\left(\frac{B}{\delta\ell^\alpha} + \alpha \log(J(T_1(\delta, \omega)))\right) + \mathcal{O}(1) \\ &\leq \frac{\alpha(1+\eta)}{C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})} \left[\log\left(\frac{Be(1+\eta)^\alpha}{\delta\ell^\alpha C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})^\alpha}\right) + \log\log\left(\frac{B(1+\eta)^\alpha}{\delta\ell^\alpha C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})^\alpha}\right) \right] + \mathcal{O}(1). \end{aligned}$$

Therefore, for every $\omega \in \mathcal{E}_T$, we have:

$$\begin{aligned} J(T_0(\delta, \omega), \omega) &\leq J(F'(\eta)) + J(T_1(\delta, \omega)) \\ &\leq \underbrace{\frac{\alpha(1+\eta)}{C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})} \left[\log\left(\frac{Be(1+\eta)^\alpha}{\delta\ell^\alpha C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})^\alpha}\right) + \log\log\left(\frac{B(1+\eta)^\alpha}{\delta\ell^\alpha C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})^\alpha}\right) \right]}_{\overline{C}(\eta, \epsilon)} + \mathcal{O}(1). \end{aligned}$$

Therefore, $\mathcal{E}_T \subset \{J(\tau_\delta) \leq \overline{C}(\eta, \epsilon) + T_\epsilon\}$ for every $T \geq \overline{C}(\eta, \epsilon) + T_\epsilon$. By definition of expectation, we have:

$$\begin{aligned} \mathbb{E}[J(\tau_\delta)] &= \int_{\Omega} J(\tau_\delta, \omega) d\mathbb{P} \leq \int_{J(\tau_\delta) \leq \overline{C}(\eta, \epsilon) + T_\epsilon} J(\tau_\delta, \omega) d\mathbb{P} + \int_{J(\tau_\delta) > \overline{C}(\eta, \epsilon) + T_\epsilon} J(\tau_\delta, \omega) d\mathbb{P} \\ &\leq T_\epsilon + \overline{C}(\eta, \epsilon) + \sum_{T=T_\epsilon + \overline{C}(\eta, \epsilon)}^{\infty} \mathbb{P}(J(\tau_\delta) > T) \\ &\leq T_\epsilon + \overline{C}(\eta, \epsilon) + \sum_{T=1}^{\infty} \mathbb{P}(\mathcal{E}_T^c) \end{aligned}$$

Then by Lemma 7, there exists two constants B and C such that we have:

$$\sum_{T=1}^{\infty} \mathbb{P}(\mathcal{E}_T^c) \leq \sum_{T=1}^{\infty} BT \exp(-CT^{1/8}).$$

So we have a bound that:

$$\mathbb{E}[J(\tau_\delta)] \leq T_\epsilon + \frac{\alpha(1+\eta)}{C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})} \left[\log\left(\frac{Be(1+\eta)^\alpha}{\delta\ell^\alpha (C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c}))^\alpha}\right) + \log\log\left(\frac{B(1+\eta)^\alpha}{\delta\ell^\alpha (C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c}))^\alpha}\right) \right] + \sum_{T=1}^{\infty} BT \exp(-CT^{1/8}) + \mathcal{O}(1).$$

Thus, by dividing $\log(1/\delta)$ and let δ decreases to 0, we have:

$$\liminf_{\delta \rightarrow 0} \frac{\mathbb{E}_\mu[J(\tau_\delta)]}{\log(1/\delta)} \leq \frac{\alpha(1+\eta)}{C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})}$$

Letting η and ϵ go to zero, it is easy to see that $C_\epsilon^*(\boldsymbol{\mu}, \mathbf{c})$ converges to $T^*(\boldsymbol{\mu})$ we have that

$$\liminf_{\delta \rightarrow 0} \frac{\mathbb{E}_\mu[J(\tau_\delta)]}{\log(1/\delta)} \leq \alpha T^*(\boldsymbol{\mu}).$$

□

Proof of Lemma 7. By a union bound:

$$\mathbb{P}(\mathcal{E}_T^c) \leq \sum_{t=h(T)}^T \mathbb{P}(\hat{\boldsymbol{\mu}}(t) \notin \mathcal{I}_\epsilon) + \sum_{t=h(T)}^T \mathbb{P}(\hat{\mathbf{c}}(t) \notin \mathcal{J}_\epsilon)$$

From (Garivier & Kaufmann, 2016, Lemma 19), there exists B and C such that

$$\sum_{t=h(T)}^T \mathbb{P}(\hat{\boldsymbol{\mu}}(t) \notin \mathcal{I}_\epsilon) \leq BT \exp(-CT^{1/8}).$$

So it is sufficient to bound the second term concerning the concentration of costs. We have:

$$\sum_{t=h(T)}^T \mathbb{P}(\hat{\mathbf{c}}(t) \notin \mathcal{J}_\epsilon) = \sum_{t=h(T)}^T \sum_{a=1}^K [\mathbb{P}(\hat{c}_a(t) \leq c_a - \beta) + \mathbb{P}(\hat{c}_a(t) \geq c_a + \beta)].$$

Let T be large enough such that $h(T) \geq K^2$. For any $t \geq h(T)$, we have $N_a(t) \geq \sqrt{t} - K$ for every arm a . Let $\hat{c}_{a,s}$ be the empirical mean of the first s costs from arm a such that $\hat{c}_a(t) = \hat{c}_{a,N_a(t)}$. With a union bound, we have:

$$\begin{aligned} \mathbb{P}(\hat{c}_a(t) \leq c_a - \beta) &= \mathbb{P}(\hat{c}_a(t) \leq c_a - \beta, N_a(t) \geq \sqrt{t} - K) \leq \sum_{s=\sqrt{t}-K}^t \mathbb{P}(\hat{c}_{a,s} \leq c_a - \beta) \\ &\leq \sum_{s=\sqrt{t}-K}^t \exp(-s\beta^2) \leq \frac{\exp(-(\sqrt{t}-K)\beta^2)}{1 - \exp(-\beta^2)}, \end{aligned}$$

where the second last inequality uses Hoeffding's inequality. With a same argument, we can show that the same upper bound applies to $\mathbb{P}(\hat{c}_a(t) \geq c_a + \beta)$. So we can plug in and have:

$$\begin{aligned} \sum_{t=h(T)}^T \mathbb{P}(\hat{\mathbf{c}}(t) \notin \mathcal{J}_\epsilon) &\leq 2 \sum_{t=h(T)}^T \sum_{a=1}^K \frac{\exp(-(\sqrt{t}-K)\beta^2)}{1 - \exp(-\beta^2)} \leq \frac{2KT \exp(K\beta^2)}{1 - \exp(-\beta^2)} \exp(-\sqrt{h(T)}\beta^2) \\ &\leq \frac{2KT \exp(K\beta^2)}{1 - \exp(-\beta^2)} \exp(-T^{1/8}\beta^2). \end{aligned}$$

Finally, we re-define the constants B and C to be:

$$\begin{aligned} B &\leftarrow 2 \max\{B, \frac{2KT \exp(K\beta^2)}{1 - \exp(-\beta^2)}\}, \\ C &\leftarrow \min\{C, \beta^2\}. \end{aligned}$$

So we have:

$$\mathbb{P}(\mathcal{E}_T^c) \leq \sum_{t=h(T)}^T \mathbb{P}(\hat{\boldsymbol{\mu}}(t) \notin \mathcal{I}_\epsilon) + \sum_{t=h(T)}^T \mathbb{P}(\hat{\mathbf{c}}(t) \notin \mathcal{J}_\epsilon) \leq BT \exp(-CT^{1/8}).$$

□

H Asymptotic Cost Optimality for Chernoff-Overlap in Two-armed Gaussian Bandits

To get meaningful bounds related to the lower bound, we will consider the special cases when the lower bound is known, specifically the two-armed Gaussian bandit model. In order to prove Theorem 3, we require a lemma similar to 5 which studies the convergence of empirical cost proportion. Therefore, we present Lemma 8.

Lemma 8. *Under the Chernoff-Overlap sampling rule, we have that*

$$\mathbb{P} \left(\lim_{t \rightarrow \infty} \frac{\hat{c}_a(t) N_a(t)}{J(t)} = \frac{\sqrt{c_a}}{\sqrt{c_1} + \sqrt{c_2}}, \forall a \in \mathcal{A} \right) = 1$$

Now, we can prove Theorem. 3 following the similar argument as the proof of Theorem. 6.

Proof of Theorem. 3. Let \mathcal{E} be the event that all concentrations regarding reward, cost, and empirical proportion holds, i.e.,

$$\mathcal{E} = \left\{ \forall a \in \mathcal{A}, \frac{c_a N_a(t)}{J(t)} \xrightarrow{t \rightarrow \infty} w_a^* \right\} \cap \left\{ \hat{\boldsymbol{\mu}}(t) \xrightarrow{t \rightarrow \infty} \boldsymbol{\mu} \right\} \cap \left\{ \hat{\mathbf{c}}(t) \xrightarrow{t \rightarrow \infty} \mathbf{c} \right\}$$

By Lemma. 8 and the law of large numbers, \mathcal{E} is of probability 1. On \mathcal{E} , there exists t_0 such that for all $t \geq t_0$, $\hat{\mu}_1(t) > \hat{\mu}_2(t)$ due to the concentration of empirical reward, and thus the Chernoff stopping statistics can be re-written as:

$$\begin{aligned} Z_2(t) &= N_1(t) d(\hat{\mu}_1(t), \hat{\mu}_{1,2}(t)) + N_2(t) d(\hat{\mu}_2(t), \hat{\mu}_{1,2}(t)) \\ &= J(t) \cdot \left[\left(\frac{N_1(t)}{J(t)} + \frac{N_2(t)}{J(t)} \right) I_{\frac{N_1(t)/J(t)}{N_1(t)/J(t) + N_2(t)/J(t)}}(\hat{\mu}_1(t), \hat{\mu}_2(t)) \right]. \end{aligned}$$

By Lemma 3, for all $a \geq 2$, the mapping $(\mathbf{w}, \boldsymbol{\lambda}, \mathbf{c}) \rightarrow \left(\frac{w_1}{c_1} + \frac{w_2}{c_2} \right) I_{\frac{w_1/c_1}{(w_1/c_1 + w_2/c_2)}}(\lambda_1, \lambda_2)$ is continuous at $(w^*(\boldsymbol{\mu}), \boldsymbol{\mu}, \mathbf{c})$. Therefore, for all $\epsilon > 0$ there exists $t_1 \geq t_0$ such that for all $t \geq t_1$ and all $a \in \{2, \dots, K\}$,

$$\left(\frac{N_1(t)}{C(t)} + \frac{N_2(t)}{C(t)} \right) I_{\frac{N_1(t)/C(t)}{N_1(t)/C(t) + N_2(t)/C(t)}}(\hat{\mu}_1(t), \hat{\mu}_2(t)) \geq \frac{w_1^*/c_1 + w_2^*/c_2}{1 + \epsilon} I_{\frac{w_1^*/c_1}{w_1^*/c_1 + w_2^*/c_2}}(\mu_1, \mu_2)$$

Hence, for any $t \geq t_1$, we have:

$$Z_2(t) \geq J(t) \frac{w_1^*/c_1 + w_2^*/c_2}{1 + \epsilon} I_{\frac{w_1^*/c_1}{w_1^*/c_1 + w_2^*/c_2}}(\mu_1, \mu_2) = \frac{J(t)}{(1 + \epsilon) T^*(\boldsymbol{\mu})}.$$

Notice that the algorithm will end no later than arm 2 is eliminated. Consequently, we can bound the cumulative cost a stopping time τ_δ as follows:

$$\begin{aligned} J(\tau_\delta) &\leq J \left(\inf \{ t \in \mathbb{N} : Z_2(t) \geq \beta(t, \delta) \} \right) \leq J(t_1) \vee J \left(\inf \{ t \in \mathbb{N} : J(t)(1 + \epsilon)^{-1} T^*(\boldsymbol{\mu})^{-1} \geq \log(Bt^\alpha / \delta) \} \right) \\ &\leq J(t_1) \vee (1 + \epsilon) T^*(\boldsymbol{\mu}) \left(\log \left(\frac{B}{\ell^\alpha \delta} \right) + \alpha \log(J(\tau_\delta)) \right) + \mathcal{O}(1), \end{aligned}$$

for some positive constant B from Proposition 2. Using Lemma 6, it follows that on \mathcal{E} , for $\alpha \in [1, e/2]$

$$J(\tau_\delta) \leq J(t_1) \vee \alpha(1 + \epsilon) T^*(\boldsymbol{\mu}) \left[\log \left(\frac{Be((1 + \epsilon) T^*(\boldsymbol{\mu}))^\alpha}{\ell^\alpha \cdot \delta} \right) + \log \log \left(\frac{B((1 + \epsilon) T^*(\boldsymbol{\mu}))^\alpha}{\ell^\alpha \cdot \delta} \right) \right] + \mathcal{O}(1).$$

Thus we have:

$$\limsup_{\delta \rightarrow 0} \frac{J(\tau_\delta)}{\log(1/\delta)} \leq (1 + \epsilon) \alpha T^*(\boldsymbol{\mu}) = \frac{2(1 + \epsilon) \alpha (\sqrt{c_1} + \sqrt{c_2})^2}{(\mu_1 - \mu_2)^2}.$$

Letting ϵ go to zero concludes the proof. \square

H.1 Proof of Lemma. 8

Proof of Lemma. 8. Under the Chernoff-Overlap sampling rule, we will pull arm 2 at time t when the following condition is satisfied:

$$\sqrt{\widehat{c}_2(t)}N_2(t) \leq \sqrt{\widehat{c}_1(t)}N_1(t).$$

This gives us the condition:

$$\left(\frac{\widehat{c}_2(t)}{\widehat{c}_1(t)}\right)^{\frac{1}{2}} N_2(t) \leq N_1(t).$$

Set $c(t) = \left(\frac{\widehat{c}_2(t)}{\widehat{c}_1(t)}\right)^{\frac{1}{2}}$. It is clear that $c(t)$ can be upper and lower bounded by constants as the cost is bounded. Then we pull arm 1 whenever $N_1(t) < c(t)N_2(t)$. From the above condition, it is clear that as $t \rightarrow \infty$ it must be the case that $N_a(t) \rightarrow \infty$ since $c(t)$ is both upper and lower bounded. Additionally, we have that

$$c(t)N_2(t) - c(t) \leq N_1(t) \leq c(t)N_2(t) + 1.$$

This gives us that

$$\frac{\widehat{c}_1(t)c(t)N_2(t)}{J(t)} - \frac{c(t)}{J(t)} < \frac{\widehat{c}_1(t)N_1(t)}{J(t)} < \frac{\widehat{c}_1(t)c(t)N_2(t)}{J(t)} + \frac{1}{J(t)}.$$

Since $\mathbf{c} \in \mathcal{C}$ which has bounded distribution, it must be the case that $c(t) \leq 1/\ell^2$ due to the boundedness assumption. Additionally, we have $J(t) \geq \ell t$. Therefore exists some t_0 such that for every $t \geq t_0$,

$$\frac{\max\{1, c(t)\}}{J(t)} < \epsilon/2.$$

It follows that for all $t \geq t_0$,

$$\left| \frac{\widehat{c}_1(t)N_1(t)}{J(t)} - \frac{\widehat{c}_1(t)c(t)N_2(t)}{J(t)} \right| < \epsilon/2.$$

Now note that can upper and lower bound $J(t)$ as follows:

$$\begin{aligned} J(t) &= \widehat{c}_1(t)N_1(t) + \widehat{c}_2(t)N_2(t) \leq \widehat{c}_1(t)(c(t)N_2(t) + 1) + \widehat{c}_2(t)N_2(t), \\ J(t) &= \widehat{c}_1(t)N_1(t) + \widehat{c}_2(t)N_2(t) \geq \widehat{c}_1(t)(c(t)N_2(t) - c(t)) + \widehat{c}_2(t)N_2(t). \end{aligned}$$

Therefore, we have:

$$\frac{\widehat{c}_1(t)c(t)N_2(t)}{\widehat{c}_1(t)(c(t)N_2(t) + 1) + \widehat{c}_2(t)N_2(t)} \leq \frac{\widehat{c}_1(t)c(t)N_2(t)}{J(t)} \leq \frac{\widehat{c}_1(t)c(t)N_2(t)}{\widehat{c}_1(t)(c(t)N_2(t) - c(t)) + \widehat{c}_2(t)N_2(t)} \quad (4)$$

Let $\mathcal{E} = \{\widehat{\mathbf{c}} \rightarrow \mathbf{c}\}$, so that $\mathbb{P}(\mathcal{E}) = 1$ by the strong law of large numbers. Then on \mathcal{E} , taking the limit of (4) as $t \rightarrow \infty$ yields

$$\frac{\sqrt{c_1}}{\sqrt{c_1} + \sqrt{c_2}} \leq \lim_{t \rightarrow \infty} \frac{\widehat{c}_1 c N_2(t)}{J(t)} \leq \frac{\sqrt{c_1}}{\sqrt{c_1} + \sqrt{c_2}}$$

This gives us that there exists some $t_1 \geq t_0$ such that for every $t \geq t_1$

$$\left| \frac{\widehat{c}_1 c N_2(t)}{J(t)} - \frac{\sqrt{c_1}}{\sqrt{c_1} + \sqrt{c_2}} \right| < \epsilon/2$$

Thus, combining (H.1) and (H.1), for every $t \geq t_1$

$$\begin{aligned} \left| \frac{\widehat{c}_1 N_1(t)}{J(t)} - \frac{\sqrt{c_1}}{\sqrt{c_1} + \sqrt{c_2}} \right| &= \left| \frac{\widehat{c}_1 N_1(t)}{J(t)} - \frac{\widehat{c}_1 c N_2(t)}{J(t)} + \frac{\widehat{c}_1 c N_2(t)}{J(t)} - \frac{\sqrt{c_1}}{\sqrt{c_1} + \sqrt{c_2}} \right| \\ &\leq \left| \frac{\widehat{c}_1 N_1(t)}{J(t)} - \frac{\widehat{c}_1 c N_2(t)}{J(t)} \right| + \left| \frac{\widehat{c}_1 c N_2(t)}{J(t)} - \frac{\sqrt{c_1}}{\sqrt{c_1} + \sqrt{c_2}} \right| < \epsilon/2 + \epsilon/2 = \epsilon \end{aligned}$$

Since $\frac{\widehat{c}_2 N_2(t)}{J(t)} = 1 - \frac{\widehat{c}_1 N_1(t)}{J(t)}$, we also have that $\frac{\widehat{c}_2 N_2(t)}{J(t)} \rightarrow \frac{\sqrt{c_2}}{\sqrt{c_1} + \sqrt{c_2}}$ as desired. \square

I δ -PAC Analysis for CTAS and CO

In this section, we provide a proof for the δ -PAC guarantees of both algorithms following the identical procedure from Garivier & Kaufmann (2016) for completeness. However, we only present the proof for the case where $\alpha > 1$. The case of $\alpha = 1$ can also be generalized using the same argument as Theorem 10 in Garivier & Kaufmann (2016).

Proof of Proposition 1. The proof relies on the fact that $Z_{a,b}(t)$ can be expressed using function I_α introduced in Definition (2). An interesting property of this function, that we use below, is the following. It can be checked that if $x > y$,

$$I_\alpha(x, y) = \inf_{x' < y'} [\alpha d(x, x') + (1 - \alpha)d(y, y')].$$

For every a, b that are such that $\mu_a < \mu_b$ and $\widehat{\mu}_a(t) > \widehat{\mu}_b(t)$, one has the following inequality:

$$\begin{aligned} Z_{a,b}(t) &= (N_a(t) + N_b(t)) I \frac{N_a(t)}{N_a(t) + N_b(t)} (\widehat{\mu}_a(t), \widehat{\mu}_b(t)) \\ &= \inf_{\mu'_a < \mu'_b} N_a(t) d(\widehat{\mu}_a(t), \mu'_a) + N_b(t) d(\widehat{\mu}_b(t), \mu'_b) \\ &\leq N_a(t) d(\widehat{\mu}_a(t), \mu_a) + N_b(t) d(\widehat{\mu}_b(t), \mu_b). \end{aligned}$$

One has

$$\begin{aligned} \mathbb{P}_\mu(\tau_\delta < \infty, \widehat{a}_{\tau_\delta} \neq a^*) &\leq \mathbb{P}_\mu(\exists a \in \mathcal{A} \setminus a^*, \exists t \in \mathbb{N} : \widehat{\mu}_a(t) > \widehat{\mu}_{a^*}(t), Z_{a,a^*}(t) > \beta(t, \delta)) \\ &\leq \mathbb{P}_\mu(\exists t \in \mathbb{N} : \exists a \in \mathcal{A} \setminus a^* : N_a(t) d(\widehat{\mu}_a(t), \mu_a) + N_{a^*}(t) d(\widehat{\mu}_{a^*}(t), \mu_{a^*}) \geq \beta(t, \delta)) \\ &\leq \mathbb{P}_\mu\left(\exists t \in \mathbb{N} : \sum_{a=1}^K N_a(t) d(\widehat{\mu}_a(t), \mu_a) \geq \beta(t, \delta)\right) \\ &\leq \sum_{t=1}^{\infty} e^{K+1} \left(\frac{\beta(t, \delta)^2 \log(t)}{K} \right)^K e^{-\beta(t, \delta)}. \end{aligned}$$

The last inequality follows from a union bound and (Magureanu et al., 2014, Theorem 2), originally stated for Bernoulli distributions but whose generalization to one-parameter exponential families is straightforward. Hence, with an exploration rate of the form $\beta(t, \delta) = \log(Bt^\alpha/\delta)$, for $\alpha > 1$, choosing B satisfying

$$\sum_{t=1}^{\infty} \frac{e^{K+1}}{K^K} \frac{(\log^2(Bt^\alpha) \log t)^K}{t^\alpha} \leq B$$

yields a probability of error upper bounded by δ . \square

Proof of Proposition 2. In CO, the event where we incorrectly identify the best arm corresponds to incorrectly eliminating the best arm at some point. Therefore, we once again have

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\mu}}(\tau_{\delta} < \infty, \hat{a}_{\tau_{\delta}} \neq a^*) &\leq \mathbb{P}_{\boldsymbol{\mu}}(\exists t \in \mathbb{N} : \exists a \in \mathcal{A} \setminus a^* : N_a(t)d(\hat{\mu}_a(t), \mu_a) + N_{a^*}(t)d(\hat{\mu}_{a^*}(t), \mu_{a^*}) \geq \beta(t, \delta)) \\ &\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \sum_{a=1}^K N_a(t)d(\hat{\mu}_a(t), \mu_a) \geq \beta(t, \delta)\right) \\ &\leq \sum_{t=1}^{\infty} e^{K+1} \left(\frac{\beta(t, \delta)^2 \log(t)}{K}\right)^K e^{-\beta(t, \delta)}. \end{aligned}$$

The result then follows from the same line of reasoning as Proposition 1 afterwards. \square

Remark 1. For the proof of $\alpha = 1$, see (Garivier & Kaufmann, 2016, Theorem 10). The same proof can be used to show that both CTAS and CO are δ -PAC for $\alpha = 1$, where $B = 2K$.